Jour. Korean Earth Science Society, v. 26, no. 6, p. 511–523, August 2005

한국 지구과학 올림피아드 문항 분석을 통한 문항의 질 향상 방안

이기영[1,*] · 김찬종[2]

[1]한성과학고등학교, 120-080 서울특별시 서대문구 현저동 산 5번지
[2]서울대학교 사범대학 지구과학교육과, 151-748 서울특별시 관악구 신림동 산 56-1번지

Analysis of Korea Earth Science Olympiad Items for the Enhancement of Item Quality

Ki-Young Lee[1,*] and Chan-Jong Kim[2]

[1]Hansung Science High School, san 5, Hyeonjeo-Dong, Seodaemun-Gu, Seoul 120-080, Korea
[2]Department of Earth Science Education, Seoul National University, Seoul 151-748, Korea

Abstract: The purpose of this study is to analyze the 1st and 2nd Korea Earth Science Olympiad (KESO) items, in order to find informations to enhance item quality. To do this, internal and external item classification frameworks are developed. Item difficulty (P), discrimination index (DI), correlation, and reliability are estimated by using classical test theory. Generalizability is also estimated by applying the generalizability theory. The results of item classification are as follows: (1) "Geology", "astronomy" and "data analysis and interpretation" are dominant in content and inquiry process domain, respectively. Nearly every item has textbook context. (2) There is no difference between the preliminary and final tests in terms of their thinking skills sections. (3) As a whole, the ratio of items with pictures is high in item representation. However, multiple-choice and short answer items are more common in preliminary competition, and essay type items are found more often in final competition. The ratio of simple items is high in middle school section and preliminary competition, but composite items are dominant in high school section and final competition. The findings of item analysis are as follows: (1) In the middle school section, P is low and DI is moderate. But in the high school section, there is a considerable differences between science high schools and other high schools in general. (2) The highest correlation is reported between the scores of meteorology domain and total score in middle school, whereas in high school astronomy domain and total score show the highest correlation. (3) General high school section show the highest Cronbach α and generalizability. (4) General high school section show acceptable generalizability coefficient (> 0.80), but middle and science high school section should increase the number of items to reach acceptable generalizability level.

Keywords: item classification frameworks, item analysis, olympiad, reliability, generalizability

요 약: 본 연구에서는 한국 지구과학 올림피아드 문항의 질 향상 방안을 모색하고자 1회와 2회 예선 및 본선 문항을 다양한 측면에서 분석하였다. 문항 분석을 위해 내적 및 외적 문항 분류틀을 구안하여 적용하였다. 고전 검사이론을 적용하여 문항의 난이도와 변별도, 상관계수, 그리고 신뢰도를, 일반화가능도 이론을 적용하여 일반화가능도를 각각 추정하였다. 문항 분류틀 적용 결과는 다음과 같다: (1) 내용 차원에서는 지질 영역과 천문 영역에, 지식 및 탐구과정 차원에서는 자료 분석 및 해석에 집중되는 경향을 보였다. 또한 맥락 차원에서는 거의 대부분의 문항이 교과서적인 상황을 토대로 한 것이었다. (2) 요구 사고 수준에서 예선과 본선 간의 차이를 발견할 수 없었다. (3) 문항 표현 범주에서는 그림의 사용 비율이 가장 높았고, 문항 유형 범주에서는 예선은 선다형과 단답형의 비율이, 본선은 서술형의 비율이 높았다. 또한 문항 형식 범주에서는 중학부와 예선은 단독형의 비율이 높았고, 과고부와 본선은 복합형의 비율이 높았다. 문항을 분석한 결과는 다음과 같다: (1) 중학부는 대체로 난이도가 낮고, 변별도는 적절하였다. 그러나 과고부는 일반고와 과학고간에 상당한 차이가 있었다. (2) 중학부는 대기 영역과 총점과의 상관이, 과고부는 천문 영역과 총점과의 상관이 가장 높았다. (3) 양호한 문항이 가장 많은 일반고부의 문항 내적일관성신뢰도와 일반화가능도가 가장 높았다. (4) 일반

*Corresponding author: leeky321@chol.com
Tel: 82-17-424-8098
Fax: 82-2-363-5892

고는 출제된 문항 수로 적정 수준의 일반화가능도에 도달되나, 중학부나 과학고는 출제된 문항보다 2배 이상 많은 문항수가 필요한 것으로 나타났다.

주요어: 문항 분류틀, 문항 분석, 올림피아드, 신뢰도, 일반화가능도

서 론

연구의 필요성 및 목적

우리나라에서 과학 올림피아드의 역사는 외국에 비하면 매우 짧은 편이다. 국제 물리올림피아드(IPhO)와 국제 화학올림피아드(IChO), 그리고 국제 생물올림피아드(IBO)가 2004년에 각각 제 36회, 제 35회, 제 13회 대회를 개최한 반면, 한국물리올림피아드(KPhO)와 한국화학올림피아드(KChO)는 1999년에 시작되었으며, 한국생물올림피아드는 2004년에 제 8회 대회를 개최하였다. 지구과학 분야에서는 1996년 시작된 국제천문올림피아드(IAO)가 있으나 국제 지구과학올림피아드(IESO)는 아직 개최되지 못한 상황이다.

한국지구과학올림피아드(KESO)는 우리나라 중·고등학생들의 지구과학에 대한 관심 고양, 지구과학 학습 능력의 향상 도모, 지구과학 분야의 영재를 발굴, 그리고 차기 국제 지구과학 올림피아드 대회에 참가할 국가대표 학생 후보자 양성을 위한 목적으로 2003년 처음으로 실시되었다(한국지구과학회, 2003, 2004). 올해로 3회를 맞이하는 KESO는 중학부, 일반고부, 과학고부로 구분하여 실시하며, 1회와 2회 대회는 전국 16개 지역 예선을 거쳐 전국 본선 대회를 개최하였으나, 3회 대회부터는 지역 예선과 본선 구분없이 1차로 주·객관식 문항을 이용한 지필평가를 전국 단위로 실시하고, 1차 평가 결과로 참가 학생수의 30% 내외의 학생을 선발한 후 2차로 심층 문답식으로 실습·수행평가를 실시한다.

KESO에서 지역 예선과 전국 본선 문항은 지구과학을 전공한 교수와 교사에 의해 개발되었으며, 지구과학 전 영역에 걸쳐 출제되었다. 하지만 2회를 거쳐오는 동안 출제된 문항에 대한 분석이 제대로 실시되지 못하였던 것이 현실이다. 그러므로 3회를 맞이하는 현 시점에서 지금까지 출제된 문항들에 대한 면밀한 분석이 있어야 할 것이며, 이에 따른 반성(reflection)과 피드백(feedback)이 뒤따라야 할 것으로 본다.

문항 분석(item analysis)이라 함은 문항을 작성할 때 개개 문항이 제대로 그 기능을 수행하고 있는 지를 확인하는 작업으로, 문항 양호도 분석이라고도 한다(변창진 외, 2001). 본 연구에서는 다양한 측면에서의 문항 분석을 위해 고전 검사 이론(classical test theory)을 적용하여 문항의 난이도(item difficulty), 변별도(discrimination index), 상관계수(correlation), 그리고 신뢰도(reliability)를 산출하며, 일반화가능도 이론(generalizability theory)을 적용하여 보다 향상된 개념의 신뢰도인 일반화가능도를 산출한 것이다.

고전검사이론을 통해 산출되는 신뢰도는 측정 도구의 신뢰도 즉, 측정 결과의 일관성(consistency)에 대한 추정 방법에 집중되어 왔기 때문에 측정 상황에서 발생할 수 있는 여러 오차요인에 대한 설명이 불충분하다(Brennan, 2000; Burns, 1998). 고전검사이론의 이러한 약점을 보완하여 다중오차요인(multiple sources of error)의 분산성분의 크기와 이들 간의 상호작용 효과를 동시에 추정할 수 있게 하기 위해 등장한 것이 바로 일반화가능도 이론이다(Cronbach et al., 1972). 일반화가능도 이론은 분산분석(ANOVA) 체제를 적용하여 측정상황에서 발생할 수 있는 다중 오차 요인을 동시에 분석하고, 측정점수에 대한 오차요인의 상대적 영향력을 산출하여 일반화가능도 계수와 함께 의사 결정자에게 안정적인 점수를 얻기 위한 측정조건을 제시함으로써 신뢰도 추정 과정을 한 단계 향상시킨 것이다(김성숙과 김양분, 2001). 본 연구에서는 이러한 검사 이론들을 적용하여 1회와 2회에 출제된 KESO 문항들을 분석함으로써 다양한 측면에서 문제점들을 찾아내고, 이를 토대로 타당도와 신뢰도가 높은 양질의 문항을 개발하기 위한 방안을 모색해보고자 한다.

연구 방법

연구 대상 및 절차

본 연구는 2003년과 2004년에 각각 실시된 1회와 2회 한국지구과학올림피아드 지역 예선문항과 전국 본선 문항 총 106개를 대상으로 하였다. Table 1에서 보는 바와 같이 분석 대상 문항 중 중학부 문항은 52문항이었으며, 고교부는 54문항이었다. 2003년에는 예선과 본선, 중학부와 고교부의 문항수가 모두 같았으나, 2004년에는 예선과 본선 문항수와 중학부와 고

Table 1. Number of items analyzed

Middle school					High school					Total
Pre.		Final			Pre.		Final			
2003	2004	2003	2004	sum	2003	2004	2003	2004	sum	
12	15	12	13	52	12	16	12	14	54	106

Table 2. Internal item classification framework

차원	영역	세부 영역
내용	A. 천문	A1. 태양계 A2. 행성의 운동 A3. 별의 특성 A4. 은하의 우주
	B. 지질	B1. 지구의 구조 B2. 지각의 물질 B3. 지각변동(판구조론) B4. 지구의 이사 B5. 지질조사의 지질도
	C. 대기	C1. 일기 변화 C2. 대기 중의 물 C3. 대기의 운동 C4. 대기의 순환
	D. 해양	D1. 해저지형 D2. 해수의 성질 D3. 해류 D4. 해파 D5. 조석
	E. 지구환경	E1. 온실효과(지구온난화) E2. 엘니뇨 E3. 물의 순환
지식 및 탐구과정	I. 지식	I1. 기억 I2. 이해 I3. 적용
	II. 문제인식 및 가설설정	II1. 문제 인식 II2. 가설 설정
	III. 탐구설계 및 수행	III1. 변인 설정 III2. 실험 장치 고안 및 제작 III3. 관찰, 측정, 분류, 실험 III4. 실험 절차
	IV. 자료분석 및 해석	IV1. 정량적 분석 IV2. 정성적 분석 IV3. 자료의 변환
	V. 결론도출 및 일반화	V1. 결론 도출 V2. 일반화
맥락	a. 교과서적	문항의 자료나 상황이 교과서를 토대로 한 경우.
	b. 탈교과서적	문항의 자료나 상황이 교과서를 토대로 하지 않은 경우(STS적인 상황 등)

교육의 문항수가 각각 달랐다. 문항 분석을 하기 위해 우선 선행 연구를 토대로 2개의 문항 분류틀을 구안하였으며, 이 분류틀을 사용하여 올림피아드 문항을 시행 연도와 예선과 본선으로 구분하여 분석하였다. 또한 2004년 서울 지역의 올림피아드 예선 문항 채점 결과를 이용하여 문항의 난이도와 변별도, 상관계수, 그리고 신뢰도를 산출하였다.

문항 분석을 위한 분류틀 구안

본 연구에서는 지구과학 올림피아드 문항 분석을 위해 우종옥 외(1996)와 홍미영 외(2002)의 연구에서 사용한 문항 분류틀을 수정하여 내적문항 분류틀과 외적 문항 분류틀로 구분하여 2개의 문항 분류틀을 구안하였다.

1) 내적 문항 분류틀(internal item classification framework)

내적 문항 분류틀은 문항이 포함하는 내적인 요소들을 분석하기 위한 것으로 Table 2에서 같이 내용(content), 지식 및 탐구과정(knowledge & Inquiry process), 맥락(context)의 3차원으로 구안되었다. 내용 차원(dimension)은 다시 지구과학의 내용 영역(domain)에 따라 천문, 지질, 대기, 해양, 지구환경으로 분류하였으며 각 영역들은 몇 개의 단원(unit)으로 세분하였다. 지식 및 탐구과정 차원은 4개의 일반적 탐구과정 영역(문제인식 및 가설설정, 탐구설계 및 수행, 자료분석 및 해석, 결론도출 및 일반화)에 지식 영역을 추가하여 하나의 차원으로 만들었으며, 각 영역들은 몇 개의 요소들로 나누었다. 맥락 차원은 자세하게 나누지 않고 문항의 맥락이 교과서 상황이냐 아니냐에 따라 크게 "교과서적"과 "탈교과서적" 2개의 영역으로만 구분하였다.

2) 외적 문항 분류틀(external item classification framework)

외적 문항 분류틀은 문항을 구성하는 외적 요소들을 분석하기 위한 것으로 Table 3에서 같이 문항 표현(item representation), 문항 유형(item type), 문항 형식(item form)의 3개 범주(category)로 구분하여 구안되었다. 문항 표현 범주는 무엇을 사용하여 발문을 구성하였느냐에 따라 그림, 그래프, 표, 보기로 구분하였으며, 문항 유형 범주는 피험자들이 답을 하는 방식에 따라 선다형, 단답형, 서술형의 3개 유형으로 구분하였다. 문항 형식은 하나 문항의 존재 여부에 따라 단독형과 복합형으로 구분하였다.

문항 분류틀 적용

구안된 2개의 분류틀을 2003년과 2004년 예선 및

Table 3. External item classification framework

대범주	소범주	세부 설명
문항 표현	그림	지문이 사진이나 삽화 등의 '그림으로' 제시된 경우.
	그래프	지문이 x와 y축을 가진 '그래프로' 제시된 경우.
	표	지문이 '표로' 제시된 경우.
	보기	지문에 대한 설명이나 해석이 <보기>로 제시된 경우.
문항 유형	선다형	여러 개의 답지(option) 중 1개 또는 그 이상을 고르게 하는 경우.
	단답형	단답의 형태로 답만 쓰게 한 경우.
	서술형	답에 대한 설명 또는 이유, 풀이과정을 서술하게 한 경우.
문항 형식	단독형	하부 문항 없이 문항이 하나로만 구성된 경우.
	복합형	문항이 2개 이상의 하부 문항으로 구성된 경우.

본선 106문항에 대해 각각 적용시켰다. <부록 1>은 2004년 중학교부 예선 15문항을 대상으로 내적 문항 분류틀을 적용시킨 예를 보여준다. 3개의 차원에 따라 각 문항이 해당되는 세부영역을 기록하였으며, 하나의 문항은 각 차원의 영역 중 하나에만 해당되도록 하여 2개의 칸에 중복되는 경우는 없도록 분석하였다. <부록 2>는 외적 문항 분류틀을 적용시킨 예이다. 이 분류틀의 경우는 특성상 하나의 문항이 여러 칸에 중복될 수 있도록 분석하였다. 예를 들어, 어떤 문항의 발문이 그림과 표를 사용하여 표현되었다면 두 칸에 모두 표시를 하며, 문항 형식이 복합형일 경우는 하부 문항들의 유형이 서로 다를 수 있으므로 이 또한 중복 표시가 가능하도록 하였다.

문항의 난이도와 변별도 및 상관계수 산출

본 연구에서는 학생 개개인의 채점된 답안지를 구할 수 있었던 2004년도 서울 지역 예선 결과를 예시로 문항 분석을 실시하였다. 분석 대상은 중학부 참가자 72명과 고교부 참가자 40명의 문항별 점수 자료였다. Table 4는 분석 대상의 기술통계치이다.

1) 난이도와 변별도

본 연구 대상 문항 대부분이 부분 점수를 가진 서술형 문항이므로 문항의 난이도(item difficulty: P)는 선다형 문항 난이도 공식을 수정하여 사용하였다.

$$P = \frac{R}{N \times SA} \times 100$$

R: 어떤 한 문항에서 전체 응답자들이 받은 점수의 합
SA: 그 문항에 주어진 배점
N: 총 사례수

Table 4. Descriptive statistics of subject

	Middle school	High school	
		General	Science
N of Cases	72	30	10
Mean	70.0	33.4	52.0
SD	14.0	17.7	12.0

난이도에 의하여 문항을 평가하는 절대적인 기준은 없으나 본 연구에서는 Cangelosi(1990)의 기준을 적용하였다. 난이도 .25 이하는 "어려운 문항"으로, .25 ~ .75는 "적절한 문항"으로, .75 이상은 "쉬운 문항"으로 판단하였다.

문항의 변별도는 변별도 지수(discrimination index: DI)를 사용하였으며, SPSS 프로그램의 Correlation을 이용하여 각 문항과 총점의 상관계수를 구하였다. 변별도 또한 평가의 절대적인 기준은 없으나 본 연구에서는 Ebel(1965)의 기준을 적용하였다. .40 이상은 변별도가 "높은 문항"으로, .30~.39는 "있는 문항"으로, .20~.29는 "낮은 문항"으로, .10~.19는 "매우 낮은 문항"으로, .10 미만은 "없는 문항"으로 판단하였다.

2) 상관계수

상관계수에 따른 상관관계를 언어적으로 표현하는 절대적인 기준은 없으나 본 연구에서 상관관계의 판단은 성태제(2002)의 기준을 적용하였다. 상관계수 .00 이상 .20 미만은 "상관이 거의 없다"로, .20 이상 .40 미만은 "상관이 낮다"로, .40 이상 .60 미만은 "상관이 있다"로, .60 이상 .80 미만은 "상관이 높다"로, .80 이상 1.00 미만은 "상관이 매우 높다"로 판단하였다.

신뢰도 산출

1) 고전검사이론의 신뢰도

고전검사이론에서 사용하는 여러 가지 신뢰도 중 본 연구에서는 문항의 내적일관성신뢰도(internal consistency reliability)인 Cronbach α를 산출하였다.

2) 일반화가능도

일반화가능도 이론은 크게 일반화 연구(generalizability study, G 연구)와 결정 연구(decision study, D 연구)로 나뉘어진다. G 연구는 과학 수행평가 점수에 어떤 오차 요인이 얼마만큼 영향을 주는지 그 상대적 크기를 분석하기 위해 실시한다. 우선 오차요인에 따라 국면(facet)을 설정하고 자료 수집 형태가 교차(crossed)모형인지 내재(nested)모형인지 결정하여 분산분석(ANOVA) 설계를 적용한다. 그 다음은 분산분석 결과 얻어진 각 분산원의 제곱평균(MS)으로부터 분산성분(variance component)을 추정하여, 분산성분의 상대적 크기를 비교하여 각 오차원의 영향력을 분석한다(이기영, 2004). 본 연구에서 G 연구는 문항에 대한 문항(i)을 단일 국면으로 하는 교차 설계인 p×i 설계를 사용하였는데 모든 피험자(p)의 모든 문항(i)을 1명의 채점자가 채점한 것으로 가정한다.

D 연구는 과학 수행평가 점수가 얼마나 신뢰로운지 그리고 신뢰로운 평가가 되기 위해서는어떤 조건을 갖추어야 하는지 알아내기 위하여 실시한다. G 연구에서 산출된 오차원의 분산성분을 토대로 고전검사이론의 신뢰도 계수와 유사한 개념인 일반화가능도 계수를 산출한다. 또한 오차분산의 각 국면의 수를 늘림으로써 적정 수준(0.80)의 일반화가능도 계수를 산출하기 위한 최적의 조건을 제시한다.

연구 분석의 기본적인 자료 처리는 GENOVA (GENeralized analysis Of VAriance) 프로그램을 사용하였다. GENOVA는 일반화가능도 이론을 적용시키기 위해 개발되었으며, 다른 통계 프로그램에서는 계산되지 않는 분산성분의 추정치와 비율, 일반화가능도 계수, 각 국면의 조건 변화에 따른 일반화가능도 계수의 변화와 같은 다양하고 상세한 결과를 제공한다(Crick & Brennan, 1983).

연구 결과 및 논의

내적 문항 분류틀 적용 결과

Table 5는 내적 문항 분류틀을 적용한 결과를 나타낸 것이다. 각각의 차원(dimension)별로 분류 결과를 분석하면 다음과 같다.

1) 내용 차원

분석된 전체 106문항 중 38%(41문항)가 지질 영역으로 가장 많이 출제되었으며, 그 다음은 31%(33문항)인 천문 영역, 대기가 15%(16문항), 해양이 13%

Table 5. Result of item classification by applying internal framework

Dimension	Domain	Middle school(%)					High school(%)					Total
		Preliminary		Final			Preliminary		Final			
		2003	2004	2003	2004	sum	2003	2004	2003	2004	sum	
Content	A. Astronomy	25	33	25	31	29	33	38	25	36	33	31
	B. Geology	42	40	33	38	38	42	38	33	36	37	38
	C. Meteorology	17	13	25	15	17	8	0	33	14	13	15
	D. Oceanography	17	7	17	15	13	17	13	8	14	13	13
	E. Earth environment	0	7	0	0	2	0	13	0	0	4	3
Knowledge & Inquiry process	I. Knowledge	50	7	67	15	33	58	13	0	21	22	27
	II. Identifying problems & formulating hypothesis	8	0	0	0	2	0	0	0	0	0	1
	III. Planning & performing inquiry	8	13	0	15	10	8	0	0	0	2	6
	IV. Analyzing & interpreting data	25	80	25	69	52	33	75	100	71	70	61
	V. Making & testing conclusion	8	0	8	0	4	0	13	0	7	6	5
Context	a. Textbook	92	87	83	85	87	75	100	100	100	94	91
	b. Non-textbook	8	13	17	15	13	25	0	0	0	6	9

(14문항), 지구환경이 3%(3문항)이었다.

중학교 52문항 중 가장 많은 비율을 차지한 것은 지질 영역(38%)이었으며, 그 다음은 천문 영역(29%)이었다. 대기와 해양 영역의 출제 비율은 17%와 13%로 상대적으로 낮았으며, 지구환경 영역은 52문항 중 1문항이었다. 중학교에서 가장 많은 비율을 차지한 지질 영역 중에서는 지각의 물질과 지구의 역사 단원에서 주로 문항이 출제된 것으로 나타났으며, 천문 영역에서는 태양계와 행성의 운동 단원의 출제 비율이 높았다. 대기 영역은 비교적 고르게 출제 되었으며, 해양 영역은 해수의 성질과 조석 단원에서만 출제되었다. 지구환경 영역에서는 물의 순환 단원에서 1문항이 출제되었다.

고등학교도 중학교의 경우와 거의 비슷하였다. 전체 54문항 중 37%가 지질 영역으로 가장 많은 비율을 차지하였고, 천문 영역이 그 다음으로 33%, 대기와 해양 영역이 각각 13%씩을 차지하였으며, 지구환경 영역은 4%를 차지하였다. 지질 영역에서는 여러 단원에서 비교적 고르게 출제되었으며, 천문 영역에서는 주로 행성의 운동 단원에서 출제된 것으로 나타났다. 대기 영역의 경우도 비교적 고르게 출제되었으나 대기의 순환 단원에서는 1문항도 출제되지 않았다. 해양 영역에서는 해수의 성질 단원이 가장 많은 비율을 차지하였으며 해저지형단원은 1문항도 출제되지 않았다. 지구환경 영역은 지구온난화와 엘니뇨 단원에서 각각 1문항씩이 출제되었다.

2003년과 2004년 모두 지질 영역이 출제 비율이 가장 높고, 그 다음으로 천문 영역인 것으로 나타났다. 예선과 본선에서 모두 2003년에 비해 2004년에 천문과 지질 영역의 출제 비율이 증가하고, 대기와 해양 영역의 비율이 감소하는 경향을 보였다. 또한 2003년에는 지구환경 영역의 출제 비율이 0%였으나, 2004년에는 중학교가 7%, 고등학교가 13%로 새롭게 추가된 것으로 나타났다. 그러나 2003년과 2004년 모두 출제 비율에서 예선과 본선간의 뚜렷한 차이를 발견할 수는 없었다. 또한 2003년에는 예선과 본선 모두 12문항으로 동일하였으나, 2004년에는 본선 문항을 예선 문항보다 2문항씩 적게 출제하였다. 이것은 본선 문항의 수를 줄이는 반면, 예선에 비해 좀 더 심화된 문항을 출제하려는 의도인 것으로 판단된다.

2) 지식 및 탐구과정 차원

전체 문항 중 탐구과정의 자료분석 및 해석 영역이 차지하는 비율이 61%로 가장 많았으며, 그 다음이 지식 영역으로 27%를 차지하였다. 가장 많은 비율을 차지한 자료 분석 및 해석 영역은 중학교(52%)보다 고등학교(70%)가 더 많았다. 또한 고등학교는 중학교에 비해 정량적 분석이 정성적 분석보다 더 많았으며, 중학교의 경우는 자료의 변환이 2%로 거의 출제되지 않은 반면, 고등학교의 경우는 17%로 정성적 분석과 같은 비율로 출제된 것으로 나타났다. 지식 영역은 중학교(33%)가 고등학교(22%)보다 더 많았다. 중학교의 경우는 지식 영역 중 이해가 가장 많았고, 고등학교는 적용이 가장 많았으며 기억은 중·고 모두 0%를 차지하였다. 탐구설계 및 수행 영역의 비율은 중학교(10%)가 고등학교(2%)에 비해 약간 높았으며, 출제된 문항은 모두 III3(관찰, 측정, 분류, 실험)에 속하였다. 그러나 변인 설정, 실험장치 고안 및 배치, 실험 절차에 관련된 문항은 전혀 출제되지 않았으며, 문제인식 및 가설설정과 결론도출 및 일반화 영역의 출제 비율 또한 각각 전체의 1%와 5%로 중·고 모두 거의 출제되지 않은 것으로 나타났다. 이와 같이 나타난 것에는 여러 가지 요인이 있을 수 있지만 문항 제작시 분류틀을 사용하지 않아 탐구과정 영역에 대한 적절한 안배를 하지 못하였기 때문에 출제자마다 가장 보편적이며 제작하기 쉬운 "자료 분석 및 해석" 영역으로 집중된 것으로 판단된다.

2004년 고등학교 본선을 제외하면 대체로 지식 영역의 출제 비율이 현저하게 감소하는 경향을 보였으며, 이와 반대로 자료 분석 및 해석 영역은 출제 비율이 대체로 증가하는 경향을 나타내었다. 2004년의 경우는 본선 문항에서 중학교와 고등학교가 모두 지식 영역이 예선에 비해 증가하고 탐구과정 영역의 비율이 감소한 것으로 나타나 시험의 성격과 문항의 사고 수준이 불일치하는 문제점이 나타났다. 또한 탐구 과정에서 문제 인식이나 가설 설정, 그리고 결론 도출 및 일반화 영역은 출제 비율이 극히 낮은 것으로 나타났으며, 전반적으로 탐구 요소들이 고르게 출제되지 못하고 자료 분석 및 해석에 집중되는 경향을 보였다.

지식 및 탐구 과정 요소로만 판단하였을 때, 예선과 본선 간의 차이는 발견할 수 없었으며 본선 문항이 예선 문항보다 좀 더 고차원 사고 기능을 요구하는 문항으로 구성되었다는 근거를 찾을 수는 없었다.

3) 맥락 차원

2003년과 2004년에 출제된 거의 대부분의 문항이

Table 6. Result of item classification by applying external framework

Category	Sub-category	Middle school (%)				High school (%)			
		Preliminary		Final		Preliminary		Final	
		2003	2004	2003	2004	2003	2004	2003	2004
Item representation	Picture	42	73	42	46	67	88	50	50
	Graph	17	7	8	15	8	0	33	21
	Table	33	27	33	54	0	6	8	14
	Example	25	20	17	23	0	13	8	21
Item type	Multiple-choice	42	40	33	23	8	19	8	14
	Short answer	42	53	42	62	58	50	58	36
	Essay	33	13	42	15	50	69	58	57
Item form	Single	67	67	42	62	42	44	25	50
	Composite	33	33	58	38	58	56	75	50

교과서적인 상황을 보대로 출제된 것으로 나타났으며, STS적인 상황 등 탈교과서적인 맥락에서 출제된 문항은 전체 106문항 중 10문항(9%)에 불과하였다.

3가지 차원을 이용한 내적 문항 분류틀을 적용한 결과, 문항들이 학교 정기고사나 수능 문항에 비해 심화된 수준으로 개발되긴 하였으나 주로 지식을 이용한 수렴적 사고를 요하는 문항이었으며 발산적 사고나 창의적 사고력을 요하는 문항은 거의 출제되지 않은 것으로 나타났다.

외적 문항 분류틀 적용 결과

Table 6은 외적 문항 분류틀을 적용한 결과를 나타낸 것이다.

문항 표현 범주에서는 그림을 사용한 비율이 가장 높았으며, 그래프를 사용한 비율이 가장 낮았다. 또한 2003년에 비해 2004년에 그림을 사용한 비율이 증가되는 경향을 나타내었다. 중학교의 경우는 표나 보기를 비교적 많이 사용하였으나, 고등학교의 경우는 표나 보기를 사용한 비율이 매우 낮았다. 문항 표현 범주에서 예선과 본선의 차이는 본선에서 그림보다는 표나 보기를 사용한 비율이 증가하였다는 것이다.

문항 유형 범주에서는 중학교가 고등학교 문항에 비해 선다형 비율이 더 큰 것으로 나타났다. 중학교의 경우는 예선에서 주로 선다형이나 단답형이 출제되었으며, 본선에서는 선다형의 비율이 감소하고 단답형과 서술형의 비율이 증가하였다. 고등학교의 경우는 예선에서 단답형과 서술형이 비교적 고르게 출제되었으나, 본선에서는 상대적으로 서술형의 비율이 높은 것으로 나타났다. 그러나 2004년의 경우는 예선보다 본선에서 오히려 단답형과 서술형의 비율이 감소하였다.

문항 형식 범주에서는 중학교의 경우 단독형이 복합형보다 더 많이 출제되었으며, 고등학교는 단독형보다는 복합형이 더 많이 출제된 것으로 나타났다. 2004년 고등학교를 제외하면, 중학교와 고등학교 모두 예선보다 본선에서 복합형이 단독형보다 더 많은 비율을 차지하였다.

문항의 난이도와 변별도 및 상관계수

1) 문항 난이도와 변별도 분석

Table 7과 8은 고전적인 문항 분석 방법을 이용하여 문항의 난이도(P)와 변별도(DI)를 산출한 결과이다.

중학교의 경우 평균 난이도가 0.69로 전반적으로

Table 7. Item difficulty (P) and discrimination index (DI) of items for middle school

No. of item	domain	allotment	P	DI
1	B	6	0.99	0.18
2	B	6	0.84	0.46
3	B	6	0.88	0.58
4	B	10	0.87	0.50
5	B	6	0.90	0.44
6	B	6	0.30	0.33
7	C	9	0.69	0.74
8	D	6	0.93	0.53
9	E	4	0.68	0.36
10	C	12	0.60	0.55
11	A	6	0.62	0.60
12	A	6	0.78	0.26
13	A	4	0.15	0.07
14	A	9	0.52	0.67
15	A	4	0.67	0.18
	Mean		0.69	0.43

Table 8. Item difficulty (P) and discrimination index (DI) of items for high school

No. of item	domain	allotment	General high		Science high	
			P	DI	P	DI
1	B	6	0.03	-0.10	0.25	0.56
2	B	3	0.22	0.39	0.27	0.73
3	B	7	0.44	0.25	0.20	-0.07
4	B	6	0.44	0.56	0.80	0.17
5	B	5	0.03	-0.11	0.10	0.23
6	B	7	0.28	0.62	0.57	0.50
7	D	7	0.40	0.59	0.47	0.61
8	E	3	0.63	0.42	0.77	-0.19
9	E	9	0.21	0.07	0.34	-0.12
10	D	7	0.48	0.49	0.59	0.27
11	A	6	0.32	0.63	0.50	0.60
12	A	5	0.40	0.80	0.70	0.54
13	A	8	0.29	0.73	0.60	0.72
14	A	6	0.72	0.58	0.83	0.12
15	A	6	0.35	0.82	0.95	0.35
16	A	6	0.34	0.74	0.70	0.56
Mean			0.35	0.47	0.54	0.35

다소 쉬운 편이었으며, 변별도는 0.43으로 양호하였다. 그러나 1, 2, 3, 4, 5, 8번 문항은 너무 쉬웠고, 13번 문항은 너무 어려웠던 것으로 나타났으며, 1, 13, 15번은 변별력이 없는 문항으로 분석되었다.

고교부의 경우는 일반고 학생들에게는 평균 난이도가 0.35로 다소 어려웠고, 과학고 학생들에게는 난이도가 0.54로 적절했던 것으로 나타났다. 또한 일반고와 과학고 모두 변별도는 양호하였으며 일반고가 과학고에 비해 조금 높은 것으로 나타났다. 일반고의 경우는 1, 5번 문항이 매우 어려웠고 부적(negative) 변별도를 보이는 좋지 않은 문항으로 분석되었다. 그러나 과학고의 경우는 13, 14번 문항이 너무 쉬웠으며 5번 문항이 가장 어려웠던 것으로 분석되었고, 3, 8, 9번이 부적 변별도를 보이는 좋지 않은 문항인 것으로 나타났다.

난이도와 변별도 분석 결과, 전반적으로 일반고와 과학고의 결과에 있어 상당한 차이가 있음을 확인할 수 있었다. 일반고 학생들에게 매우 양호한 문항이 과학고 학생들에게는 매우 좋지 않은 문항이었거나 또는 정반대인 문항들이 몇몇 발견되었다.

2) 문항 양호도 분석

Table 9-11은 Table 6과 7의 결과를 토대로 문항의 양호도를 판단한 것이다. 양호도의 판단 기준은 Cangelosi(1990)와 Ebel(1965)의 기준을 적용하여 난이도가 0.25와 0.75 사이이고 변별도가 0.30 이상인 문항을 양호한 것으로 판단하였다.

Table 9. Matrix of P×DI of middle school

DI \ P	<.25	.25-.49	.50-.75	>.75
<.10	13			
.10-.19			15	1
.20-.29				12
.30-.40		6	9	
>.40			7, 10, 11, 14	2, 3, 4, 5, 8

Table 10. Matrix of P×DI of general high school

DI \ P	<.25	.25-.49	.50-.75	>.75
<.10	1, 5, 9			
.10-.19				
.20-.29				
.30-.40	2			
>.40		4, 6, 7, 10, 11, 12, 13, 15, 16	8, 14	

중학부의 경우는 전체 15문항 중 6개 문항이 양호한 것으로 나타났다. 고교부의 경우는 일반고가 전체 16문항 중 11문항이 양호하였으며, 과학고는 8문항이 양호한 것으로 나타났다. 일반고와 과학고의 양호한 문항 중 겹치는 것은 6, 7, 11, 12, 13, 16번으로 6문항이었다.

Table 11. Matrix of P×DI of science high school

DI \ P	<.25	.25-.49	.50-.75	>.75
<.10	3	9		8
.10-.19				4,14
.20-.29	5		10	
.30-.40				15
>.40		1, 2, 7	6, 11, 12, 13, 16	

3) 상관계수 분석

Table 12~14는 내용 영역간의 상관과 총점과의 상관을 나타낸 것이다.

중학부의 경우는 5개 영역 중 총점과의 상관이 가장 높은 것은 0.80으로 대기 영역(C)이었으며, 지구환경 영역(E)이 0.36으로 가장 낮게 나타났다. 이 결과로 볼 때, 대기 영역에서 높은 점수를 얻은 학생들이 총점에서도 높은 점수를 얻는다고 할 수 있다. 영역 간에는 천문 영역(A)과 대기 영역(C)의 상관이 0.44로 가장 높았으며, 지질 영역(B)과 지구환경 영역(E)이 가장 낮았다.

고교부의 경우는 대기 영역(C)에서 출제된 문항이 없었기 때문에 4개 영역으로만 상관계수를 산출하였다. 일반고는 총점과의 상관이 가장 높은 것은 0.94로 천문 영역(A)이었으며, 지구환경 영역(E)이 0.28로 가장 낮았다. 영역 간에는 천문 영역(A)과 해양 영역(D)의 상관이 0.51로 가장 높았으며, 지질 영역(B)과 지구환경 영역(E)이 부적 상관을 나타내었다. 과학고의 경우도 일반고와 마찬가지로 4개 영역 중 총점과의 상관이 가장 높은 것은 0.78로 천문 영역(A)이었으며, 지구환경 영역(E)과는 부적 상관을 나타내었다. 영역 간에는 천문 영역(A)과 해양 영역(D)의 상관이 0.38로 가장 높았으며, 지질 영역(B)과 지구환경 영역(E) 그리고 천문 영역(A)과 지구환경 영역(E)이 부적 상관을 나타내었다. 고교부의 경우 천문 영역에서 높은 점수를 얻는 학생이 총점에서도 높은 점수를 얻는 것으로 분석되었다.

중학부와 고교부 모두 지구환경 영역이 총점과의 상관이 가장 낮거나 부적 상관을 보인 것으로 나타난 것이 특징적이라 할 수 있겠다.

검사 도구의 신뢰도

1) 고전검사이론의 신뢰도 계수

Table 15는 중학부와 고교부에 사용된 문항의 내적 일관성신뢰도인 Cronbach α를 산출한 것이다. 일반고부가 0.815로 신뢰도가 가장 높게 산출되었으며, 과학고부가 0.582로 일반고부나 중학부에 비해 낮게 산출되었다. 일반고부에서 신뢰도가 가장 높게 산출된 것은 난이도가 적절하고, 변별도가 높은 양호한 문항이 가장 많았기 때문인 것으로 분석된다.

Table 12. Correlation coefficients between content domains of middle school

	A	B	C	D	E	sum
A	-	0.29	0.44	0.27	0.30	0.74
B	0.29	-	0.40	0.37	0.04	0.74
C	0.44	0.40	-	0.37	0.39	0.80
D	0.27	0.37	0.37	-	0.10	0.53
E	0.30	0.04	0.39	0.10	-	0.36
sum	0.74	0.74	0.80	0.53	0.36	-

Table 13. Correlation coefficients between content domains of general high school

	A	B	D	E	sum
A	-	0.46	0.51	0.22	0.94
B	0.46	-	0.26	-0.11	0.68
D	0.51	0.26	-	0.38	0.66
E	0.22	-0.11	0.38	-	0.28
sum	0.94	0.68	0.66	0.28	-

Table 14. Correlation coefficients between content domains of science high school

	A	B	D	E	sum
A	-	0.01	0.38	-0.20	0.78
B	0.01	-	0.31	-0.32	0.57
D	0.38	0.31	-	0.01	0.67
E	-0.20	-0.32	0.01	-	-0.17
sum	0.78	0.57	0.67	-0.17	-

Table 15. Reliability coefficients

	Middle school	High school	
		General	Science
N of Items	15	16	16
Cronbach α	0.700	0.815	0.582

2) 일반화가능도 계수

Table 16은 일반화가능도 이론에서 G 연구 결과로 산출된 각 분산 성분의 비율을 나타낸 것이다. 표에서 p는 전집분산이며, i와 pi, e는 오차분산에 해당된다. 중학부와 고교부 모두 전체 분산 중 학생과 문항의 상호작용 성분(pi)과 오차(e)가 포함된 잔차(residual) 분산이 가장 큰 비율을 차지하는 것으로

Table 16. Variance components of p×i design

Source of Variation	Percentage of Total Variance (%)		
	Middle school	High school	
		General	Science
Persons (p)	7.3	17.9	6.5
Items (i)	35.7	20.1	34.1
pi, e	57.0	62.0	59.4

Table 17. Generalizability coefficients of p×i design

Number of Facet	G-coefficient()		
	Middle(15)*	High(16)	
		General	Science
Ni=1	0.113	0.224	0.098
Ni of G study	0.657	0.822	0.636
Level 0.80	Ni=32	Ni=14	Ni=37

*The figure of ( ) indicates the number of items used in G study

나타났다. 문항간의 차이를 의미하는 문항 분산(i)은 일반고부가 20.1%로 가장 작았고, 중학부가 35.7%로 가장 크게 산출되었다. 반면, 학생간의 차이를 의미하는 피험자분산(p)은 과학고부가 6.5%로 가장 작았고, 일반고부가 17.9%로 가장 크게 산출되었다. 문항 분산이 차지하는 비율이 작다는 것은 문항 간에 차이가 작아 문항의 특성이 학생들의 점수에 미치는 영향이 작다는 것으로 해석할 수 있으며, 피험자 분산의 비율이 크다는 것은 피험자의 특성이 점수에 미치는 영향이 크다는 것을 의미한다. 또한 잔차가 전체 분산 중 가장 크다는 것은 명확하게 설명될 수는 없지만 문항 구인 이외에 오차분산에 기여하는 또 다른 변동요인이 있을 수 있음을 암시한다.

Table 17은 위의 G 연구를 토대로 수행된 D 연구의 결과로 추정된 일반화가능도 계수를 나타낸 것이다. 일반적으로 G 연구에서 전집분산의 비율이 클수록, 오차분산의 비율이 작을수록 일반화가능도 계수는 크게 추정된다. D 연구 결과, 일반고부의 일반화가능도 계수가 가장 크게 추정되었으며 과학고부가 가장 작게 추정되었다. 또 일반고의 경우는 출제된 16개 문항보다 작은 14개 문항 정도로도 적정 수준의 일반화가능도 계수인 0.80에 도달되는 반면, 중학부의 경우는 0.80의 계수를 얻기 위해서는 32개의 문항이, 과학고부의 경우는 37개의 문항이 필요한 것으로 나타났다. Fig. 1은 문항수 증가에 따른 일반화가능도 계수의 변화를 나타낸 것이다.

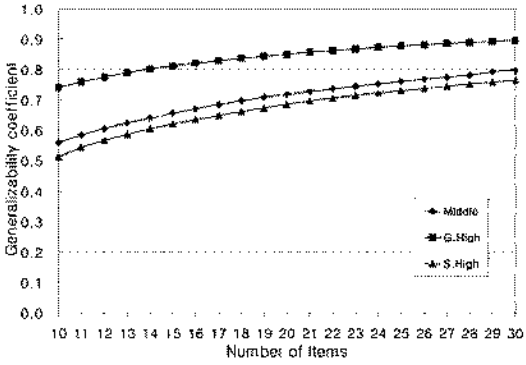

Fig. 1. Change of G-coefficient by increase of item number for p×i design.

결론 및 제언

본 연구에서는 한국 지구과학 올림피아드 문항의 질을 향상시키기 위한 방안을 모색하고자 2003년과 2004년에 실시된 제 1회와 제 2회 예선 및 본선 문항 106개를 대상으로 다양한 측면에서 문항 분석을 실시하였다.

문항 분석을 위해 선행 연구를 토대로 내적 문항 분류틀과 외적 문항 분류틀을 구안하여 이를 106개의 문항에 적용하였다. 또한 2004년 서울 지역 예선 문항 채점 결과를 이용하여 고전 검사이론을 적용하여 문항의 난이도와 변별도, 상관계수, 그리고 내적 일관성신뢰도를 산출하였으며, 일반화가능도 이론을 적용하여 한 단계 향상된 개념의 신뢰도인 일반화가능도를 추정하였다.

내적 문항 분류틀 적용 결과, 내용 차원에서는 중학부와 고교부 모두 지질 영역과 천문 영역에 집중되는 경향을 나타내었고, 상대적으로 대기와 해양 영역의 비율이 낮았으며, 지구환경 영역은 극히 낮은 비율을 차지하였다. 지식 및 탐구과정 차원에서는 전반적으로 탐구 요소들이 고르게 출제되지 못하고 자료 분석 및 해석에 집중되는 경향을 보였다. 또한 예선과 본선 간의 차이는 발견할 수 없었으며 본선 문항이 예선 문항보다 좀 더 고차원 사고 기능을 요구하는 문항으로 구성되었다는 근거를 찾을 수는 없었다. 맥락 차원에서는 거의 대부분의 문항이 교과서적인 상황을 토대로 출제된 것으로 나타났으며, STS적인 상황 등 탈교과서적인 맥락에서 출제된 문항의 비율은 매우 낮았다.

외적 문항 분류틀 적용 결과, 문항 표현 범주에서

는 그림의 사용 비율이 가장 높았고, 그래프 사용 비율이 가장 낮았다. 문항 유형 범주에서는 예선에서는 선다형과 단답형의 비율이 본선에 비해 상대적으로 높은 반면, 본선에서는 서술형의 비율이 예선에 비해 높았다. 문항 형식 범주에서는 대체로 중학부와 예선에서는 단독형의 비율이 높고, 고교부와 본선에서는 복합형의 비율이 높았다.

위 두 문항 분류틀 적용 결과를 종합해보면, 올림피아드의 성격으로 볼 때 예선과 본선 간의 내적 수준의 차이가 있어야 하나, 실제 결과에서는 뚜렷한 차이를 발견할 수 없었다. 또한 발산적 사고나 창의력을 요하는 문항은 거의 출제되지 않은 것으로 나타났다. 하지만 외적 수준에서는 본선 문항이 예선 문항에 비해 복합형이면서 서술형인 문항의 비율을 높임으로써 고차원적 사고를 요하도록 표현된 것으로 분석되었다.

한편, 2004년 서울 예선 결과를 보대로 다양한 측면에서 문항을 분석한 결과는 다음과 같다. 문항 난이도와 변별도의 경우, 중학부는 전반적으로 난이도가 낮은 편이었으며, 변별도는 적절하였다. 고교부의 경우 일반고는 난이도가 높은 편이었고, 과학고는 난이도가 적절하였다. 변별도는 일반고와 과학고 모두 양호하였다. 난이도와 변별도로 판단한 문항 양호도 분석에서는 양호한 문항이 일반고가 가장 많고 과학고가 가장 적어, 같은 문항을 사용하더라도 일반고와 과학고에서 상당한 차이를 나타내는 것을 알 수 있었다. 내용 영역과 종심 간의 상관을 분석한 결과, 중학부는 대기 영역에서 높은 점수는 얻는 것이 가장 유리하였으며, 고교부는 천문 영역에서 높은 점수를 얻는 것이 가장 유리한 것으로 나타났다.

문항 내적일관성신뢰도 분석 결과, 양호한 문항이 가장 많은 일반고부의 신뢰도가 가장 높은 것으로 나타났으며, 일반화가능도 이론을 적용한 결과 또한 이와 같았다. 문항간의 차이를 의미하는 문항분산이 가장 작고, 학생간의 차이를 의미하는 피험자분산이 가장 높은 일반고에서 일반화가능도 계수가 가장 크게 추정됨으로써 가장 신뢰도가 높은 것으로 나타났다. 또 일반고의 경우는 출제된 문항 수 정도면 적정 수준의 일반화가능도에 도달되는 반면, 중학부나 과학고의 경우는 출제된 문항보다 2배 이상 많은 문항 수가 필요한 것으로 분석되었다.

이상과 같은 연구 결과를 보대로 보다 질 높은 올림피아드 문항을 개발하기 위한 제언을 하면, 우선 올림피아드 문항 출제 시에 개별 문항 카드 이외에 본 연구에서와 같은 내적 및 외적 문항 분류틀을 이용한 전체 개발 문항에 대한 분류 작업이 선행되어야 할 것이다. 이 과정에서 내용 영역이나 탐구과정 영역이나 요소, 그리고 문항 표현 방식 등의 적절한 비율 조정이 이루어짐으로써 전반적인 선사도구의 질이 향상될 수 있을 것이다. 본 연구의 분류틀 적용 결과에서도 알 수 있듯이 2004년 예선과 본선 문항은 상당 부분 출제 비율 조정과 문항 수정이 이루어졌어야 했던 것으로 판단할 수 있다.

그 다음으로는 다양한 측면에서 문항 분석을 수행하여 양호한 문항들이 출제될 수 있도록 해야 할 것이다. 하지만 성취도 평가와는 달리 올림피아드의 성격상 예비선사(pilot test)를 실시하여 문항 분석을 한 다음 양호도가 높은 문항만을 출제한다는 것은 현실적으로 불가능하다. 그렇다고 해서 지금처럼 양호도 선증 없이 올림피아드를 치르는 것도 불합리하다. 그러므로 향후 실시될 제 3회 올림피아드에서는 적어도 1회와 2회의 문항 분석 결과를 이용한 수정에 있을 것이다. 본 연구에서와 같은 문항 분석 결과를 이용하여 양호도가 높은 문항만을 중학부, 일반고부, 과학고부 별도로 보관하여 문제은행(item pool)을 만든 다음, 이 문항들을 대회 때마다 일정 비율로 수정 또는 보완하여 사용한다거나, 이를 보대로 새로운 문항을 개발한다면 양호도가 보장된 양질의 문항만을 출제할 수 있을 것이다. 또한 같은 문항이라도 일반고부와 과학고부에서 다른 결과를 나타낼 수 있으므로 현실적으로 어려움이 있겠지만 신뢰롭고 일반화가능성이 있는 올림피아드 문항이 되기 위해서는 일반고부와 과학고부의 문항 중 일부를 학교 교과과정 등의 특성에 맞게 다르게 출제하는 방향도 고려되어야 할 것으로 본다. 더불어 본 연구의 결과가 차후 실시될 한국 지구과학올림피아드 대회에서 질 높은 문항을 개발하는데 도움이 될 수 있기를 기대한다.

참고문헌

김성수, 김양분, 2001, 일반화가능도 이론, 교육과학사.
변창진, 최진승, 문수백, 김신규, 권대훈, 2001, 교육 평가, 학지사.
성태제, 2002, 타당도와 신뢰도, 학지사.
우종옥, 이항로, 구창현, 1996, 과학 탐구 능력 평가 문항 유형 변화에 관한 종단적 연구, 한국과학교육학회지, 16 (3), 314-328.

이기영, 2004, 평가유형과 채점 방식에 따른 중·고등학교 과학 수행평가의 일반화가능도에 관한 연구. 서울대학교 박사학위논문.

한국지구과학회, 2003, 함께하는 지구과학교육, 2 (2), 10-19.

한국지구과학회, 2004, 함께하는 지구과학교육, 3 (1), 9-13.

홍미영, 신정문, 이병룡, 이양락, 2002, 대학수학능력시험 화학II 문항에 대한 학생들의 응답 분석, 한국과학교육학회지, 22 (1), 204-213.

Brennan, R. L., 2000, Performance assessments from the perspective of generalizability theory. Applied Psychological Measurement, 24 (4), 339-353.

Burns, K. J., 1998, Beyond classical reliability: Using generalizability theory to assess dependability. Research in Nursing & Health, 21, 83-90.

Cangelosi, J. S, 1990, Designing tests for evaluating student achievement. New York: Longman.

Crick, J. E., Brennan, R. L., 1983, Manual of GENOVA: A GENeralized Analysis Of VAriance System. Iowa city, IA: American College Testing Program.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N., 1972, The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: John Wiley.

Ebel, R. L, 1965, Measuring Educational Achievement. Englewood Cliffs, NJ: Prentice-Hall.

2005년 4월 15일 원고 접수
2005년 7월 5일 수정원고 접수
2005년 7월 7일 원고 채택

<부 록>

1. Example of item classification by internal framework

Category	Sub-category	Item (2004 middle school-preliminary)															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	sum
Content	A. Astronomy											A2	A4	A1	A3	A1	5
	B. Geology	B2	B2	B2	B2	B4	B4										6
	C. Meteorology							C2			C1						2
	D. Oceanography							D2								1	
	E. Earth environment									I3							1
Knowledge & Inquiry process	I. Knowledge					I3											1
	II. Identifying problems & formulating hypothesis																0
	III. Planning & performing inquiry	III3	III3														2
	IV. Analyzing & interpreting data			IV2I	V2		IV2	IV1	IV1	IV1	IV2	IV1	IV2	IV1	IV1	IV2	12
	V. Making & testing conclusion															0	
Context	a. Scientific	a	a	a	a		a	a	a	a	a	a	a	a	a		13
	b. Non-scientific					b										b	2

2. Example of item classification by external framework

Category	Sub-category	Item (2004 middle school-preliminary)															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	sum
Item representation	Picture		○	○	○	○	○			○	○	○	○	○	○		11
	Graph							○									1
	Table	○						○	○							○	4
	Example						○					○	○				3
Item type	Multiple-choice						○			○		○	○	○		○	6
	Short answer	○	○	○		○		○	○		○		1		○		8
	Essay			○	○												2
Item form	Single		○			○	○		○	○		○	○	○	○	○	10
	Composite	○		○	○			○			○						5