중등학교 과학 수행평가의 평가 유형과 채점 방식 및 신뢰도 분석

이기영 · 안희수[1]
(한성과학고등학교) · (서울대학교)[1]

Analysis of Assessment Types, Scoring Methods and Reliability of Science Performance Assessment in Middle and High School

Lee, Ki-Young · An, Hui-soo[1]
(Hansung Science High School) · (Seoul National University)[1]

ABSTRACT

In this study, we questioned what assessment types and scoring methods of science performance assessment(SPA) were being used in middle and high school, and how much these SPA scores were reliable(generalizable). To answer these questions, SPA data obtained from the seven schools were classified according to assessment type and scoring method. Based upon this classification, we analyzed the reliability by applying generalizability theory. The result, from the classification of assessment type and scoring method, showed that SPA types of the seven schools were divided into two types: paper-pencil type and task type. Paper-pencil type included answer(content)-restricted essay-type test solely. Task type has two parts: process and outcome assessment. As the results of analyzing scoring methods of the seven schools, there were two cases in the way of scoring methods: one case is scoring all essay-type items and performance tasks by one teacher, the other is scoring assigned performance tasks by two teachers. But the case of scoring assigned essay-type items or the case of cross scoring by two or more teachers were not found. The findings of the reliability analysis are as follows: (1) Effect of essay-type item to SPA score was larger than that of performance task. (2) There was remarkable difference among the seven schools' interaction effect of person and rater in scoring performance tasks. (3) Most of generalizability(reliability) coefficients of SPA for the seven schools were smaller than the acceptable generalizability coefficient(0.80). Therefore, the population of statistical parameters such as number of item, task and rater, should be increased for approaching the acceptable generalizability level.

Key words: science performance assessment, reliability, generalizability theory, assessment type, scoring method

Ⅰ. 서 론

1. 연구의 필요성 및 목적

수행평가는 고등 정신 능력을 함양하고, 전통적인 지필 평가를 대치할 수 있는 새로운 평가방법으로, 1980년대 말부터 미국과 영국을 비롯한 많은 나라들에서 학생들의 모든 특성을 평가하기 위한 노력으로 도입하고 있으며 우리나라에서도 거의 모든 교과에 광범위하게 도입되어 실시되고 있다(성태제, 1995, 1998). 특히 과학 교과는 학생

*2004.8.23(접수) 2004.11.18(1심통과) 2005.1.12(최종통과)
Corresponding author: 이기영(leeky321@chol.com)

들의 구체적 경험(hands-on experience)을 통한 학습을 강조하고 있고, 또 학생들이 그런 활동들을 통해 탐구 기능을 학습하기 때문에 수행평가의 본질을 보다 더 잘 구현할 수 있는 교과이다(Doran *et al.*, 1998; Stiggins, 1994). 그러나 현행 과학 수행평가는 기존의 평가 방식보다 더 많은 시간과 노력을 요구하며, 제반 교육적 여건이 미비하기 때문에 수행평가가 가진 많은 장점에 기초한 자발적인 필요성이 아닌 전체 평가의 30% 이상을 수행평가로 해야 한다는 의무감에서 피동적으로 시행되고 있다. 이 때문에 원래 과학 수행평가가 추구하고자하는 평가 유형과 채점 방식과는 거리가 멀어져 있는 것이 현실이다.

수행평가가 가지고 있는 또 하나의 문제는 신뢰도(reliability)이다. 수행평가는 선다형 문항보다 완성하는데 시간이 많이 소요되고 각 평가에 포함되는 과제의 수가 적기 때문에 타당도와 신뢰도를 확보하기가 어렵다고 말한다(Linn *et al.*, 1991). 또한 내용타당도(content validity)가 높다하더라도 구인타당도(construct validity)가 낮을 수 있다(Messick, 1989). 피험자가 같은 내용 영역의 서술형 문항이나 수행과제에 다르게 반응했다면 구인(construct)이 적절하지 않은 증거가 된다. 수행평가의 신뢰도가 낮다는 것은 이러한 구인타당도의 결핍을 말해준다. 지금까지의 수행평가는 방법론적 측면만을 강조하며 타당한 성적 산출에 치중한 나머지 신뢰도 문제는 뒷전이었다. 과학 교과의 수행평가는 그 평가 도구 및 수행 과제가 다양하기 때문에 평가의 신뢰도는 매우 중요한 문제이다. 최은경(2002)의 연구에 의하면 과학 수행평가에서 중등학교 학생들은 평가의 신뢰성이나 평가 기준의 타당성에 대해 만족하지 못하고 있으며, 평가 결과에 대해 신뢰하는 정도는 낮은 것으로 나타났다. 과학 수행평가는 학생들의 성취도에 30~50% 정도 반영되는 것으로 보고되고 있으나, 학교 현장 사정이나 제반 여건의 미흡으로 인해 평가의 신뢰도가 의문시된다.

일반적으로 신뢰도의 산출은 고전검사이론(classical test theory)에서의 4개 신뢰도 추정 방법(재검사 신뢰도, 동형검사 신뢰도, 반분 신뢰도, 내적 일치도)을 사용한다. 그러나 고전검사이론을 통해 산출되는 이러한 신뢰도는 측정 도구의 신뢰도 즉, 측정 결과의 일관성(consistency)에 대한 추정 방법에 집중되어 왔기 때문에 관찰대상과 관찰자, 시기, 환경 및 상황 등의 오차요인(sources of error)을 복합적으로 고려하지 못한 약점을 가지고 있다. 다시 말해, 단순히 관찰대상의 측정 결과 또는 관찰과정이 얼마나 안정적으로 일관성 있게 기록되는가에 초점을 제한하고 있기 때문에 측정 상황에서 발생할 수 있는 여러 오차요인에 대한 설명이 불충분하다는 것이다(Burns, 1998; Brennan, 2000). 고전검사이론의 이러한 약점을 보완하여 다중오차요인(multiple sources of error)의 분산 성분의 크기와 이들 간의 상호작용 효과를 동시에 추정할 수 있게 하기 위해 등장한 것이 바로 일반화가능도 이론(generalizability theory)이다(Cronbach *et al.*, 1972). 일반화가능도 이론은 분산분석(ANOVA) 체계를 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). 수행평가의 경우 자료 수집 방법이 다양할 뿐만 아니라 측정 상황에 미치는 영향 요인이 다양하기 때문에 일반화가능도 이론을 적용하여 측정 결과에 대한 일반화 정도를 산출해야 한다는 필요성이 증가하고 있다.

이러한 필요성에 근거하여 본 연구에서는 다음과 같은 연구 문제를 설정하였다.

1) 중등학교에서 과학 수행평가는 어떤 유형으로 실시되고 있는가?

2) 중등학교에서 과학 수행평가는 어떤 방식으로 채점되고 있는가?

3) 평가 유형과 채점 방식에 따른 과학 수행평가의 일반화가능도(신뢰도)는 어느 정도인가?

2. 용어 설명

일반화가능도(generalizability): generalization과 ability의 합성어. 측정 결과를 어느 정도 일반화할 수 있느냐 하는 정도. 일반화가능도 이론에서 측정의 정확도는 그 측정의 안정성이나 일관성보다는 그 측정 결과의 일반성 혹은 보편성으로 파악하는 것이 더 타당하다고 보며, 신뢰도 계수 대신 일반화가능도 계수가 측정의 정확도에 더 적절하다고 본다.

국면(facet): 허용가능한 변동요인(source of variation)을 의미하며, 분산분석에서 쓰이는 요인(factor)과 유사한 의미로 사용된다. 국면의 수는 측정대상을 제외한 변동요인의 수로 결정된다.

Ⅱ. 연구 방법

1. 연구 대상

본 연구는 Table 1에서 보는 바와 같이 서울특별시와 경기도에 소재한 중학교 4개와 고등학교 3개의 과학 수행평가 자료를 대상으로 하였다. 중학교는 경기도 소재 2개 학교와 서울 소재 2개 학교의 과학 과목에서 총 178명, 고등학교는 서울시 소재 3개 학교 10학년 과학과 11학년 지구과학 I 과목에서 총 161명의 과학 수행평가 자료를 수집하였다. 수집된 자료에는 각 학교별 과학 교과 수행평가 계획과 한 학기 동안 실시한 구체적인 수행평가 목록, 서술형문항 및 수행과제 원본과 이에 따른 채점 기준표와 채점 결과(학생별 원점수) 그리고 각 학교 담당 교사와의 면담을 통해 파악한 채점 방식이 포함된다.

2. 신뢰도 분석 방법

본 연구에서는 과학 수행평가의 신뢰도를 분석하기 위해 고전검사이론이 아닌 일반화가능도 이론을 적용하였다. 일반화가능도 이론은 크게 일반화 연구(generalizability study, G 연구)와 결정 연구(decision study, D 연구)로 나뉘어진다.

1) G 연구

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.

2) D 연구

D 연구는 과학 수행평가 점수가 얼마나 신뢰로운지 그리고 신뢰로운 평가가 되기 위해서는 어떤 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.

3) 자료 처리

연구 분석의 기본적인 자료 처리는 GENOVA(GENeralized analysis Of VAriance) 프로그램을 사용하였다. GENOVA는 Brennan(1983)에 의해 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).

3. 신뢰도 분석을 위한 연구 설계

7개 중등학교의 과학 수행평가 자료를 수집하여 이들 수행평가의 유형과 채점 방식에 맞게 G 연구를 설계하였다. D 연구 설계는 G 연구 설계와 동일하게 하였으며, 서

Table 1. Details of the seven schools

	Middle School				High School		
	B	C	H	S1	G	K	S2
Grade	7th	7th	8th	9th	10th	11th	10th
Subject matter	Science	Science	Science	Science	Science	Earth Science	Science
Number of persons	50	38	34	56	69	30	62
Location	Kyonggi	Kyonggi	Seoul	Seoul	Seoul	Seoul	Seoul

술형문항(i)과 수행과제(t) 그리고 채점자(r) 국면은 모두 임의(random)라고 규정하였다. Table 2는 오차요인(국면)과 G 연구 및 D 연구 설계를 정리한 것이다.

여기서 p×i 설계는 문항(i)을 단일 국면으로 하는 교차 설계로, 모든 피험자(p)의 모든 서술형 수행평가 문항(i)을 1명의 교사가 모두 채점하는 것이고, p×t 설계는 과제(t)를 단일 국면으로 하는 교차 설계로, 모든 피험자(p)의 모든 수행과제(t)를 1명의 교사가 모두 채점하는 것이다. p×(t : r) 설계는 과제(t)가 채점자(r) 국면에 포함되는 2국면 부분 내재 설계로, 모든 피험자(p)를 채점하되 수행과제(t)를 2명 이상의 교사가 나누어 서로 다른 과제를 채점하는 것이다.

Table 3은 7개 학교의 연구 설계에 따른 채점자 관련 정보를 정리한 것이다.

Ⅲ. 연구 결과

1. 과학 수행평가의 유형 분석

전체 7개 학교의 과학 수행평가 자료를 수집하여 그 평가 유형을 분류한 결과, Table 4와 같이 크게 지필형(paper-pencil type)과 과제형(task type)으로 나눌 수 있었다. 지필형으로는 7개 학교 모두 응답제한형 서술형검사만을 실시하고 있었고 논술형 검사를 실시하는 학교는 한 곳도 없었다. 과제형은 과정(process) 평가와 결과물(outcome) 평가로 나눌 수 있었다. 과정 평가에는 수업태도, 준비물, 노트검사가 있었으며, 결과물 평가에는 실험보고서, 과제물, 탐구(조사)보고서, 주제발표, 과학의 달 행사가 있었다.

Table 2. G study & D study designs of the seven schools

School Level	School Name	Source of Error(Facet)	G Study Design	D Study Design
Middle school	B	t	p×t	p×T
	C	i	p×i	p×I
	H	i, t, r	p×i, p×(t : r)	p×I, p×(T : R)
	S1	t, r	p×(t : r)	p×(T : R)
High school	G	t, r	p×(t : r)	p×(T : R)
	K	i, t	p×i, p×t	p×I, p×T
	S2	t, r	p×(t : r)	p×(T : R)

*D study design use capital letter I, T, R to discriminate D study facets from G study.

Table 3. Informations about rater of the seven schools

School Level	School Name	G Study Design	No. of Item or Task	No. of Rater	Sex	Major	Career (year)
Middle school	B	p×t	3	1	F	B	4
	C	p×i	5	1	F	C	4
	H	p×i	3	1	M	E	14
		p×(t : r)	4	2	M, F	E, B	14, 8
	S1	p×(t : r)	6	2	F, F	E, P	5, 7
High school	G	p×(t : r)	8	2	M, F	C, E	19, 14
	K	p×i	5	1	M	E	7
		p×t	3	1	M	E	7
	S2	p×(t : r)	4	2	M, M	E, P	13, 16

*F: Female M: Male B: Biology C: Chemistry E: Earth Science P: Physics

Table 4. Classification of SPA types for the seven schools

		Middle School	High School
Paper-pencil type		Essay-type item (answer-restricted; content-restricted)	Essay-type item (answer-restricted; content-restricted)
Task type	Process assessment	Attitude Readiness Notebook check-up	Attitude Notebook check-up
	Outcome assessment	Experiment report Homework Project presentation Investigation report Science Month's event	Experiment report Homework Project presentation

과정 평가에서 수업태도는 수업 중 학생들의 태도를 매 시간 체크하여 감점 처리하는 것이며, 준비물은 학생들의 수업 준비 상태를 점검하는 것이며, 노트검사는 정기적으로 학생들의 필기 및 프린트 관리 상태를 검사하는 것이다. 결과물 평가에서 주제 발표는 조별로 어떤 주제에 대해 일정 기간 동안 공동으로 연구하여 그 결과를 발표하는 것이며, 과학의 달 행사는 4월 과학의 달에 치뤄지는 여러 가지 행사에 참여한 학생들의 결과물을 평가하는 것이다.

2. 과학 수행평가의 채점 방식 분석

Table 5는 7개 학교에서 실시하고 있는 과학 수행평가의 유형에 따른 채점 방식을 정리한 것이다. 수행평가의 반영 비율은 최저 30%에서 최고 50%까지 나타났으며, 표집된 7개 학교 모두 과제형 수행평가를 실시하고 있었고 지필형 수행평가는 7개 학교 중 3개 학교에서만 실시하고 있었다.

B와 C 중학교 7학년의 경우는 물상과 생물을 나누지 않고 1명의 교사가 1개 반의 과학 수업을 전담하고 있었기 때문에 모든 서술형문항과 수행과제를 1명의 교사가 채점하였다. 8학년과 9학년의 경우는 물상과 생물로 나누어서 2명의 교사가 1개 반을 분담하여 수업하였으나, 물상에서만 서술형문항 평가를 실시한 H 중학교의 경우는 물상을 담당하는 1명의 교사가 자신이 가르치는 반을 전담하여 채점하다. 수행과제 평가를 실시한 H와 S1 중학교에서는 수업을 담당하는 2명의 교사가 각자 자신이 부과한 과제를 채점하여 두 점수를 합산하는 방식을 사용하였다. G와 S2 고등학교의 경우는 10학년 과학을 2명의 교사가 수업을 분담하고 있었기 때문에 채점 또한 2명의 교사가 따로 자신의 수행과제로 평가한 후 두 점수를 합산하는 방식을 채택하였으며, K 고등학교 11학년의 경우는 1명의 교사가 수업을 전담하고 있었기 때문에 서술형문항과 수행과제를 1명의 교사가 채점하였다. 그러나 표집된 7개 학교에서 2명 이상의 교사가 한 학생의 서술형문항을 나누거나 교차하여 채점하는 경우는 없었으며, 2명 이상의 교사가 한 학생의 수행과제를 나누어 채점하고는 있었지만 교차하여 채점하는 경우는 없었다.

3. 과학 수행평가의 신뢰도 분석

1) G 연구 결과

7개 학교의 과학 수행평가 유형과 채점 방식에 따라 G 연구를 수행한 결과, 각 설계에 따른 분산성분들은 다음과 같이 추정되었다. 추정된 분산성분들 중 피험자분산(p)은 고전 검사 이론에서 진점수(true score)분산에 해당되는 전집점수(universe score)분산이며, 나머지는 모두 오차분산이다. 전집점수분산에 비해 오차분산의 비율이 커지게 되면 일반화가능도(신뢰도)는 낮아지게 된다.

Table 6은 지필형 p×i 설계에 해당되는 세 학교의 분산성분을 비교한 것이다. 세 학교에서 문항분산이 차지하는 비율은 각각 15.0%, 13.5%, 9.8%로 다른 분산성분보다 낮게 나타났는데, 이는 문항 간에 차이가 크지 않아 학생들의 수행평가 점수에 문항의 특성이 미치는 영향이 작음을 의미한다. 그러나 잔차분산이 차지하는 비율이 각각 45.6%, 54.1%, 63.6%로 다른 분산성분보다 아주 높게 나

Table 5. Scoring methods of SPA for the seven schools

School Level	School Name	Ratio (%)	Type	Details	Scoring Method
Middle school	B	40	Task	Experiment report Investigation report	Scoring all tasks with one rater
	C	40	Paper-pencil	Essay-type item	Scoring all items and tasks with one rater
			Task	Experiment report Homework Project presentation Attitude	
	H	30	Paper-pencil	Essay-type item	Scoring all items with one rater Task-nested scoring within two raters
			Task	Experiment report Notebook check-up Science Month's event	
	S1	30	Task	Experiment report Notebook check-up Science Month's event	Task-nested scoring within two raters
High school	G	40	Task	Experiment report Homework	Task-nested scoring within two raters
	K	50	Paper-pencil	Essay-type item	Scoring all items and tasks with one rater
			Task	Experiment report Project presentation	
	S2	30	Task	Experiment report Attitude	Task-nested scoring within two raters

타났는데, 이것은 이론적으로 몇 가지의 해석이 가능하다. 잔차에 혼입되어 있는 피험자와 문항의 상호작용분산성분이 큰 것으로 해석할 수도 있고, 문항 국면 이외에 오차분산에 기여하는 또 다른 변동요인이 있는 것으로도 해석할 수 있으며, 이 두 가지 모두로 해석할 수도 있다. 그러나 이 잔차분산은 이 중 어떤 것인지 구별해낼 수가 없기 때문에 해석이 불가능하여 설명할 수 없는 변량으로 판단하는 것이 일반적이다.

Table 6. Comparison of variance components for p×i design(C, H, K school)

Source of Variation	Percentage of Total Variance		
	C	H	K
Persons(p)	39.4	32.4	26.5
Items(i)	15.0	13.5	9.8
pi, e	45.6	54.1	63.6

Table 7은 과제형 p×t 설계에 해당되는 두 학교의 분산성분을 비교한 것이다. 두 학교에서 과제분산이 차지하는 비율은 1.7%와 0.0%로 매우 작게 나타난 반면, 잔차분산이 차지하는 비율이 57.3%와 68.3%로 매우 크게 산출되었다. 과제분산이 차지하는 비율이 매우 작게 나타난 것은 수행과제 간에 차이가 매우 작아 수행과제의 특성이 학생들의 수행평가 점수에 미치는 영향이 거의 없는 것으로 해석할 수 있다. 잔차가 전체 분산 중 가장 크다는 것은 명확하게 설명될 수는 없지만 과제 국면 이외에 오차분산에 기여하는 또 다른 변동요인이 있을 수 있음을 암시한다.

Table 7. Comparison of variance components for p×t design (B, K school)

Source of Variation	Percentage of Total Variance	
	B	K
Persons(p)	41.0	31.7
Tasks(t)	1.7	0.0
pt, e	57.3	68.3

Table 8. Comparison of variance components for p×(t : r) design (H, S1, G, S2 school)

Source of Variation	Percentage of Total Variance			
	H	S1	G	S2
Persons(p)	14.3	50.5	30.7	14.9
Rater(r)	1.9	0.0	0.0	3.3
t:r	0.0	4.7	31.3	4.5
pr	21.3	6.9	0.0	20.5
Residual (pt:r,e)	62.5	38.0	38.0	56.8

Table 9. Comparison of G-coefficients for p×I design (C, H, K school)

Number of Facet	G-coefficient(ρ^2)		
	C(5)*	H(3)	K(5)
Ni=1	0.464	0.375	0.294
Ni of G study	0.812	0.643	0.676
Level 0.80	Ni=5	Ni=7	Ni=10

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

Table 8은 과제형 p×(t : r) 설계에 해당되는 네 학교의 분산성분을 비교한 것이다. 학교에 따라 매우 다른 결과를 나타내었다. H와 S2 학교는 피험자분산에 비해 오차분산이 매우 크게 산출되었다. 오차분산 중에는 잔차분산이 62.5%와 56.8%로 가장 크게 산출되었으며, 잔차분산을 제외한 나머지 오차분산 중에서는 피험자와 채점자의 상호작용분산이 21.3%와 20.5%로 가장 크게 산출되었다. 그러나 S1과 G 학교는 피험자분산이 오차분산에 비해 크게 산출되었다. 잔차분산을 제외한 나머지 오차분산은 비교적 작게 산출되었으나 G 학교의 경우는 채점자 내 수행과제간의 차이가 31.3%로 매우 큰 것으로 나타났다. 네 학교 모두 채점자간 차이는 0.0~3.3%로 없거나 매우 작은 것으로 나타나 채점자 특성이 학생들의 수행평가 점수에 거의 영향을 미치지 않는 것으로 나타났으며, 채점자 내 수행과제간의 차이 또한 G 고등학교를 제외하고 매우 작게 나타났다.

2) D 연구 결과

Table 9는 세 학교의 D 연구 p×I 설계의 결과로 산출된 일반화가능도 계수(이하 G 계수)를 비교한 것이다. D

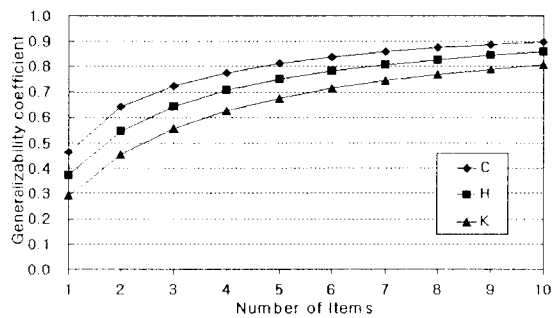

Fig. 1. Change of G-coefficient by increase of item number for p×I design (C, H, K school)

연구 결과, 국면의 기본 조건(Ni=1)에 의한 G 계수는 C 중학교가 0.464로 가장 크게 나타났으며, K 고등학교가 0.294로 가장 작게 나타났다. 이것은 교사 1명이 모든 서술형 문항을 채점하는 방식을 채택하고 있는 세 학교 중 C 중학교의 일반화가능도가 가장 높다는 것을 의미한다. 다른 두 학교에 비해 C 중학교의 일반화가능도 계수가 높게 산출된 것은 G 연구에서 피험자분산에 비해 오차분산이 더 작게 산출되었기 때문이다. C 중학교의 경우는 G

Table 10. Comparison of G-coefficients for p×T design (B, K, S2 school)

Number of Facet	G-coefficient(ρ^2)	
	B(3)*	K(3)
Nt=1	0.417	0.318
Nt of G study	0.682	0.583
Level 0.80	Nt=6	Nt=9

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

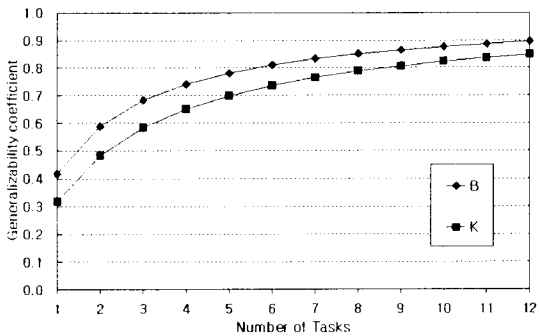

Fig. 2. Change of G-coefficient by increase of task number for p×T design (B, K school)

연구에 사용된 문항 수 정도면 일반화가능성이 있다고 판단되며, H와 K 학교의 경우는 일반화 가능한 서술형 수행평가가 되기 위해서는 더 많은 수의 문항이 필요한 것으로 판단된다. D 연구에서 산출되는 G 계수는 모든 국면의 수를 1로 하는 것을 기준으로 한다. 그러므로 국면의 수를 증가시킴으로써 G 계수를 향상시킬 수 있다. Fig. 1은 p×I 설계에서 문항 국면의 수를 증가시켰을 때 세 학교의 G 계수 변화 추이를 그래프로 나타낸 것이다. 그래프를 보면, 적정 수준의 G 계수인 0.80에 도달하기 위해서 C 중학교의 경우는 문항 5개가, H 중학교의 경우는 7개 그리고 K 고등학교의 경우는 10개가 필요한 것을 알 수 있다.

Table 10은 두 학교의 D 연구 p×T 설계의 결과로 산출된 G 계수를 비교한 것이다. D 연구 결과, 국면의 기본조건(Nt=1)에 의한 G 계수는 B 중학교가 0.417, K 고등학교가 0.318로 B 중학교의 일반화가능도가 더 높은 것으로 나타났다. 두 학교 모두 G 연구에 사용된 수행과제 수 정도로는 적정 수준의 일반화가능도에 미치지 못하며, 일반화 가능한 과제형 수행평가가 되기 위해서는 더 많은 수의 수행과제가 필요한 것으로 판단된다. Fig. 2는 p×T 설계에서 수행과제 국면의 수를 증가시켰을 때 두 학교의 G 계수 변화 추이를 그래프로 나타낸 것이다. 그래프를 보면, 적정 수준의 G 계수인 0.80에 도달하기 위해서는 B 중학교의 경우는 수행과제 6개가, K 고등학교의 경우는 9개가 필요한 것을 알 수 있다.

Table 11은 네 학교의 p×(T : R) 설계 D 연구 결과로 산출된 G 계수를 비교한 것이다. D 연구 결과, 국면의 기본조건(Nt=1, Nr=1)에 의한 G 계수는 S1 중학교가 0.530으로 가장 크게 나타났으며, H 중학교가 0.145로 가장 작게 나타났다. 이것은 교사 2명이 수행과제를 나누어 채점하는 방식을 채택하고 있는 네 학교 중 S1 중학교의 일반화가능도가 가장 높다는 것을 의미한다. 네 학교 중 S1 중학교와 G 고등학교는 G 연구에 사용된 수행과제 수 정도면 일반화가능성이 있다고 판단할 수 있으나, H 중학교와 S2 고등학교는 적정수준의 일반화가능도에 미치지 못하는 것으로 판단된다. Fig. 3은 p×(T : R) 설계에서 채점자 국면의 수를 2명으로 고정하고 수행과제 국면의 수를 증가시켰을 때 4개 학교의 G 계수 변화 양상을 나타낸 것이다. 그래프를 보면, 적정 수준의 G 계수인 0.80에 도달하기 위해서 H 중학교의 경우는 수행과제 9개와 채점자 8명 또는 수행과제 18개와 채점자 7명이 필요한 것을 알 수 있으며, S1 중학교와 G 고등학교의 경우는 모두 수행과제 3개와 채점자가 2명이, S2 고등학교의 경우는 수행과제 7개와 채점자 8명 또는 수행과제 11개와 채점자 7명이 필요한 것을 알 수 있다.

H 중학교와 S2 고등학교의 경우는 G 연구에서 피험자 분산에 비해 오차분산이 매우 크게 산출되었기 때문에 일반화가능도 계수가 낮게 추정됨으로써 수행과제와 채점자의 수가 비현실적으로 많이 필요한 것으로 나타나게 된 것이다. 또한 수행과제보다는 채점자 수를 증가시키는 것

Table 11. Comparison of G-coefficients for p×(T : R) design (H, S1, G, S2 school)

Number of Facet	G-coefficient(ρ^2)			
	H(2, 2)*	S1(2, 2)	G(4, 2)	S2(2, 2)
Nt=1, Nr=1	0.145	0.530	0.448	0.162
Nt, Nr of G study	0.352	0.796	0.866	0.379
Level 0.80	Nt=9, Nr=8 Nt=18, Nr=7	Nt=3, Nr=2	Nt=3, Nr=2	Nt=7, Nr=8 Nt=11, Nr=7

*The figure of ( ) indicates the number of (performance tasks, raters) used in G study

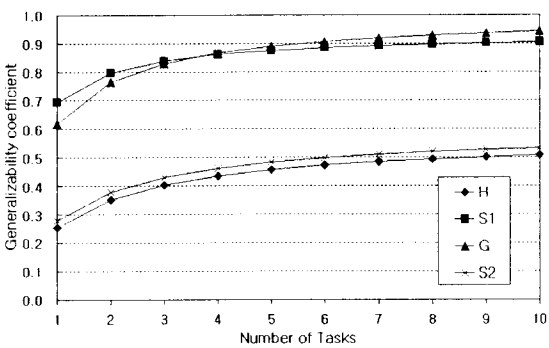

Fig. 3. Change of G-coefficient by increase of task number for p (T : R) design(H, S1, G, S2 school)

이 G 계수를 향상시키는데 더 효과적인 것으로 판단되는데, 이것은 G 연구 결과 채점자와 관련된 분산성분이 수행과제와 관련된 분산성분보다 더 크게 산출되었기 때문이다.

Ⅳ. 결론 및 제언

본 연구에서는 중등학교 과학 수행평가 자료를 수집하여 평가 유형과 채점 방식을 분석하고, 이를 토대로 지필형 및 과제형 수행평가 자료에 일반화가능도 이론을 적용하여 신뢰도(일반화가능도)를 분석하였다.

전체 7개 학교의 과학 수행평가 자료를 수집하여 그 평가 유형을 분류한 결과, 과학 수행평가의 유형이 다양하지 않으며, 과학 수행평가로 사용할 수 있는 많은 유형 중에서 일부분만을 사용하고 있는 것으로 분석되었다. 지필형 수행평가로는 과학 교과의 특성상 응답이 제한되고 내용도 제한되는 서술형 검사만을 실시하고 있었고 논술형은 사용하지 않고 있었다. 또 개방형 서술 문항보다는 단답형 서술 문항을 더 많이 사용하고 있었다. 과제형 수행평가에서는 학생의 수행 과정을 평가하는 도구로 수업태도, 준비물, 노트검사를 사용하고 있었는데, 이것은 본래 의미의 과정 평가와는 많이 동떨어진 것이었다. 결과물 평가에서는 과정 평가보다는 좀 더 다양한 종류의 과제를 포함하고 있었으나 여전히 수행평가 도입 이전에 과학 교과에서 이미 시행하고 있던 실기 평가 과제에서 크게 벗어나지 못하고 있었으며, 선진국에서 과제형 과학 수행평가로 많이 사용되는 포트폴리오, 개념도, V도 평가 등은 도입하지 못하고 있는 실정이었다.

7개 학교의 채점 방식은 거의 비슷하였다. 서술형 평가에서는 모든 학교에서 1개 반을 1명의 교사가 전담하여 수업하고 있었으며, 문항을 나누거나 교차하여 채점하지 않고 1명이 교사가 단독으로 채점하고 있었다. 과제형 수행평가의 경우도 1개 반을 1명의 교사가 전담하여 수업하는 경우는 1명의 교사가 채점하며, 2명의 교사가 1개 반을 담당하는 경우에는 각 교사가 서로 다른 과제를 실시하고, 이 결과를 각자 채점한 후 두 점수를 합산하는 방식을 사용하고 있었다.

과학 수행평가 유형과 채점 방식을 토대로 일반화가능도 이론을 적용한 결과, 서술형문항과 수행과제를 국면으로 한 1국면 교차 설계에서는 잔차성분이 매우 크게 산출되었으며, 서술형 문항과 수행과제 국면에 의한 효과가 작게 산출되었다. 또한 서술형 문항 국면에 의한 분산성분에 비해 수행과제 국면에 의한 분산성분이 매우 작게 나타나 서술형 문항 국면이 오차 분산에 더 크게 기여하는 것으로 나타났다. 이것은 두 가지로 해석이 가능하다. 우선 분석 결과 그대로 해석하여 서술형 문항의 특성이 수행과제의 특성보다 학생들의 수행평가 점수에 미치는 영향이 더 크다고 해석할 수 있을 것이다. 그러나 다른 측

면에서 보면 수행과제가 서술형 문항에 비해 채점 기준을 엄격하게 적용하지 않고 등급간의 간격을 작게 한 경우를 생각해 볼 수 있다. 원점수를 확인해 본 결과 서술형 문항에는 기본 점수가 없는 반면, 수행과제는 최하 만점의 70%에 해당되는 기본 점수를 부여하고 있었다. 수행과제와 채점자 또는 서술형 문항과 채점자를 국면으로 한 2국면 부분 내재 설계에서는 채점자간 차이가 거의 없는 것으로 나타났으며, 각 채점자들이 채점한 수행과제들간의 차이는 대체로 작은 것으로 나타났다. 이것 또한 1국면 교차 설계에서와 마찬가지로 수행과제 점수간의 간격이 작고 70%의 기본 점수를 부여하였기 때문인 것으로 판단할 수 있다. 또한 일반화가능도 계수를 산출해본 결과, 7개 학교에서 사용하는 서술형 문항이나 수행과제 수로는 대부분이 적정 수준의 일반화가능도에 도달하지 못하며, 더 많은 수의 문항과 과제가 필요한 것으로 나타났다. p×I 설계에서는 서술형 문항이 5~10개 정도가 필요한 것으로 나타났으며, p×T 설계에서는 수행과제가 6~12개 정도가 필요한 것으로 나타났다. p×(T : R) 설계에서는 두 학교의 경우는 채점자 2명에 수행과제 3개 정도면 적정 수준의 G 계수에 도달되는데 비해, 나머지 두 학교는 채점자 7~8명에 수행과제 7~18개가 필요한 것으로 나타났다.

본 연구 결과에서와 같이 서술형 문항, 수행과제, 채점자 수를 많이 늘리면 적정 수준의 신뢰도를 얻을 수는 있겠으나 이것은 현실적으로 매우 어려운 일이다. 그러므로 일반화가능도 이론에서 신뢰도의 오차가 되는 요인들의 분산을 줄일 수 있는 방안을 찾아야 할 것이다. 본 연구의 결과를 토대로 하여 그 방안을 생각해보면 다음과 같다.

먼저 서술형 문항을 이용한 수행평가에서는 문항의 내용, 난이도, 구성 형식, 채점 기준 등에서 차이를 줄이는 것이 평가의 신뢰도를 높일 수 있을 것으로 판단된다. 내용면에서 너무 동떨어진 문항들로 구성한다거나, 난이도 차이가 너무 커서 문항에 대한 학생의 반응이 일관성을 보이지 않을 경우는 신뢰도가 낮아질 수 있다. 또한 하부문항이 없는 단일 문항을 사용하거나, 채점 기준에 부분점수가 없는 경우 신뢰도가 낮아질 수 있을 것이다. 과제를 이용한 수행평가의 경우는 정량적인 연구 결과로만 보아서는 서술형 문항보다 신뢰도가 높은 평가 도구라는 판단을 내릴 수도 있겠으나, 이러한 결과는 기본 점수를 부여하여 학생간의 점수 차이가 작은 것에 기인한 것이므로 여기에 대한 정성적인 판단이 필요할 것으로 본다. 또한 2명 이상의 채점자가 과제를 나누어 채점할 경우는 채점자에 따라 많은 차이가 발생할 수 있다. 한 채점자가 다른 채점자에 비해 너무 엄격하게 채점하거나, 채점자에 따라 과제를 채점하는 기준이 많이 다른 경우는 신뢰도에 심각한 영향을 줄 수 있으므로 채점자 엄격도(severity), 채점자에 따른 채점 기준 등 채점자와 관련된 오차 분산을 줄이는 노력이 필요할 것으로 판단된다.

국 문 요 약

본 연구에서는 중등학교 과학 수행평가가 어떤 평가 유형과 채점 방식을 사용하고 있는지 분석하였으며, 이를 토대로 일반화가능도 이론을 이용하여 과학 수행평가 점수가 얼마나 신뢰로운지 분석하였다.

연구 결과, 과학 수행평가의 유형은 크게 지필형과 과제형으로 나눌 수 있었다. 지필형으로는 중등학교 모두 응답제한형 서술형검사만을 실시하고 있었다. 과제형은 과정 평가와 결과물 평가로 나눌 수 있었다. 채점 방식은 1명의 교사가 모든 서술형문항과 수행과제를 채점하거나, 2명의 교사가 수행과제를 나누어 채점하고 있었다. 그러나 2명 이상의 교사가 서술형문항을 나누거나 교차하여 채점하는 경우는 없었다.

표집된 7개 중등학교 과학 수행평가의 신뢰도 분석 결과는 다음과 같다: (1) 서술형 문항의 특성이 수행과제의 특성보다 학생들의 수행평가 점수에 미치는 영향이 더 큰 것으로 나타났다. (2) 수행과제 채점에서 채점자가 피험자를 다르게 채점하는 정도는 학교에 따라 상당한 차이가 있었다. (3) 7개 중등학교 과학 수행평가의 일반화가능도(신뢰도)는 대부분 적정 수준인 0.80에 미치지 못하는 것으로 나타났으며, 적정 수준의 일반화가능도를 얻기 위해서는 지금보다 더 많은 수의 서술형 문항과 수행과제 그리고 채점자가 필요한 것으로 분석되었다.

주요어: 과학 수행평가, 신뢰도, 일반화가능도 이론, 평가 유형, 채점 방식

참 고 문 헌

김성숙, 김양분(2001). 일반화가능도 이론. 교육과학사.

성태제(1995). 고등정신능력 신장을 위한 교육평가 방안 탐색. 국립교육평가원: 전국교육평가 심포지움 보고

서, 12, 45-90.

성태제(1998). 교육평가 방법의 변화와 결과타당도. 한국교육평가학회: 21세기 한국교육평가의 과제와 전망, 125147.

최은경(2002). 과학과 수행평가에 관한 중등학생의 인식 및 자아효능감 조사. 서울대학교 석사학위논문.

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.

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*. John Wiley: New York.

Doran, R., Chan, F., & Tamir, P.(1998). *Science educator's guide to assessment*. National Science Teachers Association, Virginia.

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

Messick, S.(1989). Validity. In R. L. Linn(Ed), *Educational Measurement*(3rd ed). American Council on Educational and MacMillan: New York, 13-103.

Stiggins, R. J.(1994). *Student-Centered Classroom Assessment*. Macmillan: New York.