교육과정평가연구
The Journal of Curriculum and Evaluation
2008, Vol. 11, No. 1, pp. 141~159

주관식 시험을 포함하고 있는 A국가고사의 채점자 효과에 대한 일반화가능도 연구

김 창 환(고려대학교 박사과정)
이 규 민(연 세 대 학 교 교 수)

≪ 요 약 ≫

일반화가능도 이론은 고전검사 이론에서 설명할 수 없었던 다중 오차원을 분석하여 검사점수의 변동요인의 상대적 영향력을 비교할 수 있는 분석 틀을 제공하며, 특히 주관식 채점 상황에서 채점자에 기인된 검사점수의 영향 정도를 파악할 수 있다는 장점이 있다. 이 연구에서는 객관식과 주관식 시험을 포함하고 있는 A국가고사를 선택하여, 주관식 시험의 채점자 요인이 검사점수의 일반화가능도에 미치는 영향을 분석하였다. 분석 결과 공통적으로 나타난 현상은 채점자 및 채점자와 관련된 효과가 거의 0에 가깝게 나타났다는 점으로 채점자를 포함하는 다른 선행 연구와는 상당히 다른 결과를 나타낸 것이다. 극히 작게 나타난 채점자 효과가 정선된 채점 기준표와 구조화된 채점자 훈련의 효과 때문이라면 현재의 측정구조를 유지하는 것이 타당하지만, 그렇지 않았다면 이를 극복할 수 있는 더 적합한 채점 체계에 대한 후속 연구가 이루어져야 할 것이다.

주제어 : 일반화가능도 이론, 다중 오차원, 주관식 시험, 채점자 관련 신뢰도

Ⅰ. 서론

A국가고사의 시험 실시 기관은 전국 16개 시·도교육청이며, 시·도교육청에서는 공동관리위원회를 두어 전국적인 규모의 시험을 관리하고 있다. 이 시험의 1차 전형은 객관식과 주관식이 부과된다.

A국가고사의 전형 요소 중 1차의 '객관식' 시험과 '주관식' 시험은 전국 공통기준을 적용하여 시행하고 있다. 이 시험의 가장 큰 특징은 한 번에 출제·채점해야 하는 분야가 현재의 국가고사 중 가장 많다는 점이다.

이 연구는 A국가고사의 1차 전형 자료 중 '주관식' 시험의 채점에 관한 것으로, 신뢰도 측면에서 현재의 채점 방식을 검토하고, 객관적이며 합리적인 채점 절차 구성을 위한 시사점을 도출하려는 목적을 갖고 진행되었다. A국가고사의 채점 원칙은 다음과 같다.

- 하나의 문항을 3인이 채점한다.
- 동일한 지역 문답지는 동일한 채점자가 채점한다.
- 문답지는 원형을 그대로 보존한다.
- 채점위원은 출제에 참여한 출제위원과 지역에서 추천한 중등 인사로 구성한다.

한편, 출제기관에서는 채점 시행 시 채점자 사이의 일관성을 높이기 위해 채점자 훈련을 시행하고 있다. 현재 출제기관에서 사용하고 있는 방식은 가채점을 활용하여 채점자 간 신뢰도를 높이는 방식으로 외국에서도 채점자 훈련과정에 일반적으로 사용되는 방식이다.

평가자(채점자)는 A국가고사의 '주관식' 시험 같은 주관식 평가 결과에 영향을 끼치는 주된 오차 요인이고, 이러한 평가자 요인이 검사점수의 신뢰도에 미치는 영향을 평가하는 것은 매우 중요하다(성태제, 2002). A국가고사는 2002학년도를 기점으로 시험의 출제 및 시행에서 많은 변화를 겪었는데 우선 출제위원 및 채점위원의 수가 대폭 확대되었으며, 출제체제 변경과 함께 채점 방식의 많은 변화를 가져왔다. 이러한 외형상의 변화에도 불구하고 채점 자체에 대한 과학적인 분석은 미진하여 국가적으로도 중요한 시험임에도 불구하고 채점자 간의 오차 요인에 대한 연구가 전혀 이루어지지 않은 상태이다. 따라서 이 연구에서는 평가자(채점자)가 포함된 측정 구조에서 각 변인이 검사점수에 미치는 상대적 영향력을 파악하고 더 효율적이고 합리적인 채점 절차 구성을 위하여 다양한 일반화가능도 연구를 수행하여 보았다.

주관식 평가 결과에 대한 측정학적 분석을 위해 일반화가능도 이론이 적합한 개념적 틀과 방법을 제공할 수 있다(김성숙, 1995; 남명호, 1996; 이규민, 2002a, 2002b, 2004; Brennan, 1992, 2001; Lee, Brennan & Frisbie, 2000; Lee & Frisbie, 1999; Lee, 2002). 전통적 평가와 달리 '평가자'를 포함하고 있는 주관식 평가에서는 일반화가능도 이론을 이용한 연구가 필수적인 요소로 인정되고 있고, 연구 결과를 활용한 다양한 측정 모형을 적용한 연구가 진행되고 있다(이영식・신상근, 2004; Gao, Brennan & Shavelson, 1994; Gao & Colton, 1996; Gao, Shavelson & Baxter, 1994; Ruiz-Primo, Baxter & Shavelson, 1993; Shavelson, Baxter & Gao, 1993; Shavelson, Gao & Baxter, 1996).

일반화가능도 이론은 고전검사 이론에 비해 오차 점수의 다양한 원천을 구별해 내고, 각각의 원천이 갖는 상대적인 영향력을 판별해 낼 수 있는 개념적인 틀과 방법론으로 정의된다(Brennan, 2001). 고전검사 이론이 관찰 점수를 진점수(true score)와 오차점수(error score)로만 구분하는 반면, 일반화가능도 이론은 오차점수에 기여하는 다양한 원천을 구분할 수 있다(이

규민, 2003; Brennan, 1992, 2001; Cronbach et al., 1972). 다양하게 수행되는 D연구(decision study)는 주어진 상황에서 가장 효율적인 측정 절차를 결정할 수 있도록 정보를 제공한다. 반면, G연구(generalizability study)는 결과가 얼마나 일반화될 수 있는가에 관심을 갖고, 모형에 따른 분산 성분을 추정하는 과정을 포함한다(김성숙・김양분, 2001; 한국교육평가학회, 2004).

Brennan(2001)이 제시한 일반화가능도 이론의 주요한 개념들을 고전검사 이론의 개념들과 관련지어 설명하면, 먼저 전집점수 분산(universe score variance)이란 용어는 고전검사 이론의 진점수 분산(true score variance)과 비슷한 개념으로 생각할 수 있다. 그러나 오차점수에 있어서는 고전검사 이론과 다른 정의를 내리고 있는데, 일반화가능도 이론은 두 가지 서로 다른 종류의 오차점수를 구분한다. 하나는 상대오차이고, 다른 하나는 절대오차인데, 상대오차는 검사 결과를 상대평가의 목적으로 활용할 경우에, 절대오차는 검사 결과를 절대평가의 목적으로 활용할 경우에 사용될 수 있다. 고전검사 이론과 달리 일반화가능도 이론에서는 상대평가와 절대평가에 적용되는 신뢰도 추정이 달라짐을 구별할 수 있다는 장점이 있다.

Ⅱ. 연구 방법

1. 연구 자료

2005학년도 A국가고사 '주관식' 문항에 대한 채점 결과의 일반화가능도 분석을 위해 이 연구에서 사용된 연구 자료는 다음의 두 종류로 구분된다. 〈표 1〉의 대상 과목은 주관식 시험 과목 중 대표적인 5개 과목을 선정하였으며 〈표 2〉의 지역은 전국 16개 시・도 지역을 대상으로 하였다.

〈표 1〉 이 연구에 사용된 연구 자료의 과목별 피험자 수와 문항 수

서울지역 5개 과목					
과목	국어	영어	수학	공통과학	공통사회
피험자 수	986명	392명	443명	163명	259명
문항 수	31문항	28문항	22문항	26문항	29문항
배점	2~4점	2~8점	2~5점	2~4점	2~5점
총점	80점	80점	80점	80점	80점

〈표 2〉 이 연구에 사용된 연구 자료의 지역별 피험자 수와 문항 수

16개 시·도 국어 과목*					
지역	서울	부산	대구	인천	광주
피험자 수	986명	476명	327명	842명	227명
지역	대전	울산	경기	강원	충북
피험자 수	470명	477명	2,283명	563명	192명
지역	충남	전북	전남	경북	경남
피험자 수	29명	406명	289명	677명	442명
지역	제주				
피험자 수	80명				

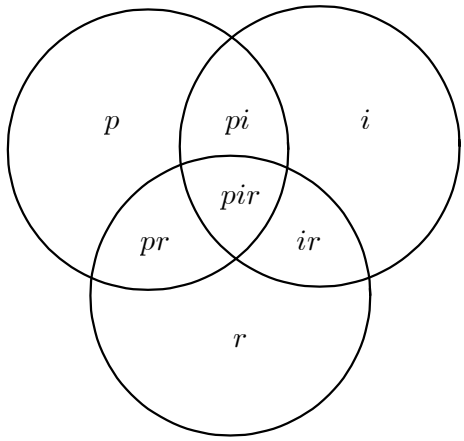
* 16개 시·도 국어 과목의 문항은 각 시·도별로 동일한 31문항임.

2. 일반화가능도 연구 모형

이 연구에서는 2005년 A국가고사 '주관식' 문항에 대한 서울지역 5개 과목과 16개 시·도별 국어 과목의 채점에 대해 일반화가능도 연구 모형을 적용하여 채점자 효과를 분석하였다. 이를 개념화하는 일반화가능도 이론의 모형은 일차원 $p \times i \times r$ 임의효과 모형이 적합하고, 선형모형은 공식 (1)과 같이 제시된다(이종성, 1988). 이 모형에서는 동일한 문항(i)으로 시험을 치른 피험자(p)를 각 평가자(r)가 채점하는 완전 교차 모형으로, 피험자(p)를 측정 대상(object of measurement)으로, 문항(i)과 평가자(r)를 임의효과 국면(facet)으로 설정하고 있다. 이러한 측정 구조로 채점이 이루어졌을 때, 검사점수를 구성하는 효과의 선형모형으로 나타내면 다음과 같다.

$$
\begin{aligned}
X_{pir} = {} & \mu && \text{전체 평균} \\
& + (\mu_p - \mu) && \text{피험자 효과} \\
& + (\mu_i - \mu) && \text{문항 효과} \\
& + (\mu_r - \mu) && \text{채점자 효과} \\
& + (\mu_{pi} - \mu_p - \mu_i + \mu) && \text{피험자와 문항의 상호작용 효과} \\
& + (\mu_{pr} - \mu_p - \mu_r + \mu) && \text{피험자와 채점자의 상호작용 효과} \\
& + (\mu_{ir} - \mu_i - \mu_r + \mu) && \text{문항과 채점자의 상호작용 효과} \\
& + (X_{pir} - \mu_{pi} - \mu_{pr} - \mu_{ir} + \mu_p + \mu_i + \mu_r - \mu) && \text{잔차 효과}
\end{aligned}
\qquad (1)
$$

선형모형을 도식화하면 〔그림 1〕과 같이 표현된다.



〔그림 1〕 G연구 $p \times i \times r$ 설계의 벤다이어그램

G연구를 통해 모형에서 제시되는 점수 효과에 의한 분산 성분이 추정되고 오차원의 상대적 영향력이 분석된다. 전체 평가점수의 분산은 각 점수 효과의 분산 성분의 합으로 구성되며, 이를 식으로 나타내면 다음과 같다.

$$\sigma^2(X) = \sigma^2(p) + \sigma^2(i) + \sigma^2(r) + \sigma^2(pi) + \sigma^2(pr) + \sigma^2(ir) + \sigma^2(pir) \qquad (2)$$

$\sigma^2(X)$는 검사점수의 분산을 나타내고, $\sigma^2(p)$는 측정의 대상인 피험자 분산, $\sigma^2(i)$는 문항 분산, $\sigma^2(r)$은 채점자 분산을 나타낸다. $\sigma^2(pi)$는 피험자와 문항 간의 상호작용 분산, $\sigma^2(pr)$은 피험자와 채점자 간의 상호작용 분산, $\sigma^2(ir)$은 문항과 채점자 간의 상호작용 분산, $\sigma^2(pir)$은 모형에서 설명되지 않는 잔차를 포함하는 분산이다.

일반화가능도 연구는 크게 두 단계로 이루어진다(Shavelson & Webb, 1994; Brennan, 2001). 첫 단계는 주어진 시험점수에 영향을 미친 요인들을 먼저 밝혀내고, 각 요인이 어느 정도의 영향을 미쳤는지를 밝혀내는 G연구(Generalizability study) 단계이다. 일반화연구 단계에서는 피험자의 검사점수에 영향을 미쳤으리라고 판단되는 요인들을 분석한 다음 각 요인들이 검사점수의 분산에 미친 영향의 크기를 알아보는 분산 성분(variance component) 추정이 이루어진다. 두 번째 단계는 일반화연구에서 추정한 분산 성분을 바탕으로 전집점수 분산, 상대오차, 절대오차 점수 분산 및 일반화가능도 계수(generalizability coefficients)와 의존도 계수

(dependability or phi coefficients)를 구하는 D연구 단계이다. D연구는 주어진 상황에서 가장 효율적인 측정 절차를 결정할 수 있도록 정보를 제공하는 것이다. 즉 각 국면(facet)의 조건을 변화시키면서 만족스러운 수준의 일반화가능도 계수가 산출되는 효율적인 측정 조건을 분석하여 낸다.

이 연구에서 식 [1]에서 설정된 일반화가능도 모형을 통해 산출된 분산 성분을 활용한 일반화가능도 계수($E\rho^2$) 와 의존도 계수(Φ)는 다음의 공식을 통해 계산된다.

$$E\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(pI) + \sigma^2(pR) + \sigma^2(pIR)} \quad (3)$$

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(I) + \sigma^2(R) + \sigma^2(pI) + \sigma^2(pR) + \sigma^2(IR) + \sigma^2(pIR)} \quad (4)$$

A국가고사와 같이 국가적으로 중요한 시험에서 일반화가능도 이론을 적용한 연구를 통해 주관식 전공시험의 채점자 효과를 분석하는 작업은 그 중요성이 인정되고, 다른 국가시험에 방법론적인 측면에서 시사점을 제공할 것으로 기대된다.

3. 자료 분석 방법

A국가고사의 일반적인 검사 특성을 파악하기 위해, 이 연구에서는 고전검사 이론을 이용한 일반적인 문항 분석을 실시하였고, 전공(주관식) 문항에 대한 기술통계와 함께 문항난이도, 변별도가 산출되었다. 검사점수의 신뢰도는 일차적으로 Cronbach α를 이용한 내적일관성 신뢰도가 산출되었다.

일반화가능도 이론을 적용하기 위해 $p \times i \times r$ 설계를 사용하였고, 임의효과 모형의 분산 성분을 추정하기 위하여 GENOVA 컴퓨터 프로그램을 사용하였다. G연구의 연구 결과에 터해 D연구는 $p \times I \times R$의 설계를 적용하여, 주관식 평가점수의 신뢰도 평가를 위해 일반화가능도 계수와 의존도 계수를 추정하였다. 채점자 및 문항 수를 변화시켜 적정 수준의 일반화가능도 수준을 확보하기 위한 효율적인 측정구조를 탐색하였다. 동일한 설계와 절차가 서울지역 5개 과목 각각에 적용되었고, 혹 지역 간에 있을 수 있는 차이를 보기 위해 16개 시·도 모두에 대해 국어 과목 대상으로 일반화가능도 연구를 진행하였다. 타 과목의 지역 간 차이는 국어 과목에 비하여 특별한 경향성을 보이지 않으므로 별도로 제시하지 않았다.

일반화가능도 연구를 위해 적용된 자료의 구조를 설명하기 위해 〈표 3〉과 같이 서울지역 국어 과목의 채점 자료를 제시하였다. 즉 총 986명의 피험자를 대상으로 동일한 3명의 채점자가 31개 문항에 대한 피험자 응답을 모두 채점한 자료 형태이다.

〈표 3〉 서울지역 국어 과목의 $p \times i \times r$ (피험자 × 문항 × 채점자) 설계 채점자료 예시

	X_{pir}											
채점자(r)	r_1				r_2				r_3			
피험자(p) 문항(i)	i_1	i_2	⋯	i_{31}	i_1	i_2	⋯	i_{31}	i_1	i_2	⋯	i_{31}
1	1	1		1	1	1		1	1	1		1
2	1	0		0	1	0		0	1	0		0
3	2	0		0	2	1		0	2	0		0
4	0	0		0	0	1		0	0	0		0
⋮												
984	0	0		0	0	0		0	0	0		0
985	2	2		1	2	2		1	2	2		1
986	1	1		1	1	1		1	1	1		1

Ⅲ. 연구 결과

1. 검사의 일반적 특성

서울지역 지원자의 국어 과목 31개 문항의 원점수상 분포를 보면 평균 28.4점, 표준편차 8.73점이고, 내적일관성 정도를 추정한 Cronbach α 신뢰도 계수는 0.79였다. 영어 과목의 경우, 원점수상 평균이 43.6점, 표준편차 13.33점으로 나타나, 다른 과목에 비해 약간 쉬웠으나 분산도는 약간 큰 것으로 나타났고, Cronbach α 계수는 0.85였다. 수학 과목은 원점수상 평균이 28.4점, 표준편차 15.95점, 신뢰도가 0.88로 나타나 가장 큰 분산도를 보였다. 공통과학과 공통사회의 원점수상 평균은 각각 25.0, 36.4이며, 표준편차는 13.52, 11.53으로 나타나 공통사회가 쉬웠던 반면 분산도는 작음을 알 수 있다. 신뢰도는 모든 교과에서 0.79 또는 그 이상의 수준으로 나타났다.

〈표 4〉 2005학년도 서울지역 A국가고사 '주관식' 전공 시험의 기술통계량

과목명	사례 수	문항 수	평균	표준 편차	최소값	최대값	신뢰도
국어	986	31	28.4	8.73	4.00	53.00	.791
영어	392	28	43.6	13.33	0.00	67.00	.850
수학	443	22	28.4	15.95	0.00	62.67	.876
공통과학	163	26	25.0	13.52	2.00	57.00	.877
공통사회	259	29	36.3	11.53	6.00	58.33	.845

서울지역 5개 교과에 대한 문항 난이도는 국어 과목의 경우, 0.2~0.4 사이의 난이도를 보인 문항이 55%로 다소 어려웠으며, 영어는 국어에 비해 난이도 측면에서 더 넓은 분포를 보여 0.5~0.8 사이의 난이도를 갖는 문항이 전체의 50% 정도를 차지했다. 수학 과목의 경우는 0.3~0.7 사이의 난이도를 가지는 문항이 전체의 60%를 차지했고 공통과학의 경우 0.2 미만과 0.2~0.4 사이의 난이도를 보인 문항이 62%로 이는 공통과학의 평균이 5개 과목 중 가장 낮은 것과 무관하지 않다. 공통사회 문항들의 난이도는 영어와 마찬가지로 그 분포가 고른 편으로 0.5를 기준으로 약 50%의 문항들이 분포하고 있다.

문항변별도 분석 결과는 문항난이도 분석과 비슷한 해석이 가능한 것으로 나타났다.

2. 일반화가능도 분석

가. 서울지역 5개 교과 분석

1) G연구 결과

〈표 5〉는 서울지역 주요 5개 과목을 대상으로 $p \times i \times r$ 설계에 대하여 피험자 분산(측정대상인 p의 분산)과 각 오차점수 분산 성분(p를 제외한 나머지 국면과 관련된 분산)의 종류와 상대적 크기를 보여주고 있다. 비교·종합하여 각 요인이 관찰점수 분산에 미치는 영향력을 설명하고 있다.

표에서 p 효과 분산 성분 추정치는 피험자 분산으로 피험자 점수가 서로 다른 정도를 나타내는 분산 성분이며, 고전검사 이론의 진점수 분산과 유사한 의미를 지닌다. r 효과 분산 성분 추정치는 채점자 분산으로 채점자가 피험자에게 서로 다른 점수를 주기 때문에 나타나는 분산이다. 예를 들어, 점수를 후하게 주는 채점자와 박하게 주는 채점자로 구성되었다면 이 채점자 분산 성분 추정치가 크게 나타날 것이고, 반면에 피험자에게 비슷한 정도로 점수를 주는 채점자들로 구성되었다면 채점자 분산이 작게 나타날 것이다. i 효과 분산 성분 추

정치는 문항의 난이도가 서로 다르기 때문에 나타나는 분산 성분이 된다. 즉 문항 간 난이도 차이가 클수록 문항 분산은 크게 나타나게 되고, 반면 문항 간 난이도 차이가 작을수록 문항 분산은 작게 나타난다. 이 밖에 $p \times r$, $p \times i$, $r \times i$는 각각 피험자와 채점자 간의 상호작용, 피험자와 문항 간의 상호작용, 채점자와 문항 간의 상호작용을 의미하게 된다. $p \times r$ 분산 성분은 같은 피험자라도 채점자에 따라 점수의 다르게 나타나는 정도라고 할 수 있으며, $p \times i$는 같은 피험자라고 하더라도 문항에 따라 점수가 다르게 나타나는 정도, $r \times i$는 같은 채점자라고 하더라도 문항에 따라 달리 점수를 부여하는 정도라고 설명될 수 있다. 마지막으로 $p \times i \times r$ 분산 성분은 피험자, 문항, 채점자의 삼원 상호작용과 모형에서 구성하고 있는 국면들로 설명되지 않은 나머지 잔차 분산을 나타내며 일반적으로는 가장 크게 나타나는 경향이 있다(이규민, 2003, 2004; Brennan, 2001).

〈표 5〉에서 국어 과목의 $p \times i \times r$ 설계에 대한 분산 성분 추정 결과를 살펴보면, 전체 분산은 0.865로 나타났고, p 효과 분산 성분 추정치 0.063은 전체 분산의 약 7.2%를 피험자 차이가 설명하고 있다고 할 수 있다. 반면 채점자의 차이 때문에 나타나는 r, $p \times r$, $r \times i$ 효과의 분산 성분은 거의 0에 가까우므로 채점자는 점수 분산에 크게 영향을 주고 있지 않음을 알 수 있다. 그러나 문항이 다르기 때문에 나타나는 i 효과와 관련 상호작용인 $p \times i$ 효과의 분산 성분은 각각 0.253, 0.497로 전체 분산 성분의 29.2%와 57.4%를 차지하여 점수 분산의 대부분이 문항 효과와 피험자와 문항의 상호작용 효과에 기인한 것으로 나타났다. 마지막으로 $p \times i \times r$ 분산 성분은 명세화한 피험자, 문항, 채점자 효과와 이들의 상호작용 효과로 설명되지 않는 잔차 분산으로 0.051이며, 전체 분산의 5.9%에 해당하여 상대적으로 그 효과가 적음을 알 수 있다. 영어, 수학, 공통과학, 공통사회 과목의 $p \times i \times r$ 설계에 대한 분산 성분 추정치도 유사하게 해석될 수 있다.

〈표 5〉 서울지역 5개 과목의 $p \times i \times r$ 설계에 대한 분산 성분 추정치와 비율*

효과	국어	영어	수학	공통과학	공통사회
p	0.063(7.2)	0.193(9.3)	0.461(20.0)	0.237(17.0)	0.133(11.5)
r	0.000(0.0)	0.000(0.0)	0.000(0.0)	0.000(0.0)	0.000(0.0)
i	0.253(29.2)	0.914(44.2)	0.397(17.3)	0.280(20.1)	0.309(26.5)
$p \times r$	0.000(0.0)	0.000(0.0)	0.000(0.0)	0.000(0.0)	0.000(0.0)
$p \times i$	0.497(57.4)	0.951(46.0)	1.429(62.1)	0.863(62.0)	0.705(60.6)
$r \times i$	0.002(0.2)	0.000(0.0)	0.000(0.0)	0.000(0.0)	0.000(0.0)
$p \times i \times r$	0.051(5.9)	0.010(0.5)	0.014(0.6)	0.011(0.8)	0.016(1.4)
계	0.865(100.0)	2.067(100.0)	2.301(100.0)	1.391(100.0)	1.164(100.0)

*( ) 안의 수치는 각 분산 성분 추정치가 전체 검사점수 분산에서 차지하는 비율임.

다만 주목할 만한 것은 국어 과목의 경우 채점자와 문항 간의 상호작용 분산[$\widehat{\sigma^2}(ri)$ =0.002] 및 잔차 분산[$\widehat{\sigma^2}(pir)$=0.050]은 각각 0.2%와 5.9%로 다른 과목에 비하여 큰 것을 확인할 수 있다. 이와 같은 결과는 다른 과목에 비해 국어 과목을 채점한 채점자들이 문항에 따라 달리 채점한 정도가 좀 크다는 것을 의미하며, 또한 피험자, 채점자, 문항의 각 국면들과 상호작용만으로 설명할 수 있는 분산 성분이 약간 작다는 것을 의미한다.

2) D연구 결과

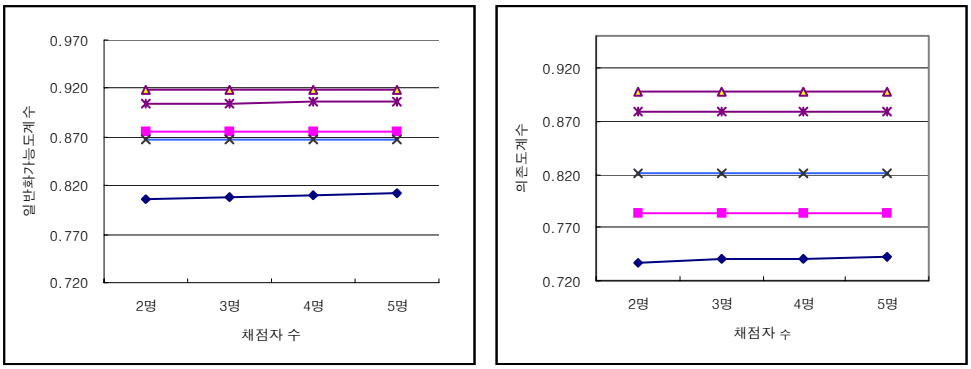
채점자와 문항 수를 조정하며 일반화가능도 계수에 미치는 영향을 알아보기 위해 다양한 D연구가 수행되었다. 채점자 수를 원래 자료가 수집될 때 채점자 수인 3명 외에 2명, 4명, 5명일 경우로 구성하였고, 문항 수는 20, 25, 30, 35개로 변화시키며 분석하였다. 각각의 D연구 결과는 〈표 6〉에 제시되어 있다

〈표 6〉 서울지역 5개 과목의 $p \times I \times R$ 설계에 대한 D연구 결과

채점자	문항 수	국어		영어		수학		공통과학		공통사회	
		G계수	D계수	G계수	D계수	G계수	D계수	G계수	D계수	G계수	D계수
2	20	0.705	0.617	0.801	0.673	0.865	0.834	0.789	0.723	0.845	0.805
	35	0.806	0.737	0.876	0.783	0.918	0.898	0.868	0.820	0.905	0.878
3	20	0.708	0.620	0.801	0.673	0.865	0.834	0.790	0.724	0.845	0.805
	35	0.809	0.740	0.876	0.783	0.918	0.898	0.868	0.821	0.905	0.879
4	20	0.710	0.621	0.802	0.674	0.865	0.834	0.79	0.724	0.846	0.805
	35	0.811	0.741	0.876	0.783	0.918	0.898	0.868	0.821	0.906	0.879
5	20	0.711	0.622	0.802	0.674	0.865	0.834	0.790	0.724	0.846	0.805
	35	0.812	0.742	0.876	0.783	0.918	0.898	0.868	0.821	0.906	0.879

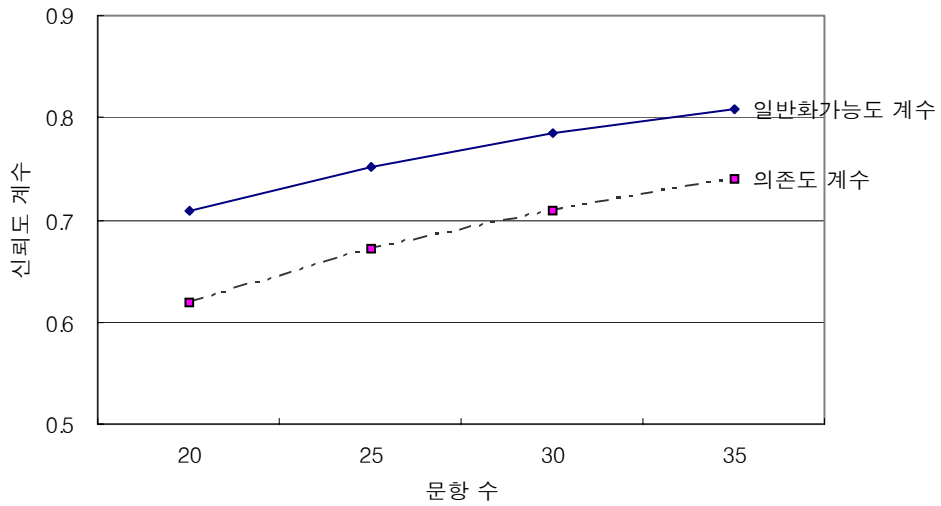
※ 1) G계수 : 일반화가능도 계수, D계수 : 의존도 계수.
2) 문항 수 변화는 20문항과 35문항으로만 표시함.

모든 과목에서 일반화가능도 계수와 의존도 계수는 채점자와 문항 수와 관계없이 0.70을 넘는 수준으로 나타났는데, 〔그림 2〕에서 볼 수 있듯이 채점자를 2명에서 3명, 4명, 5명으로 변화시켰을 때 두 계수에 큰 변화를 보이지 않았다. 다만, 국어 과목에서는 채점자 수의 증가에 따라 계수가 약간 높아지는 것을 볼 수 있다.



〔그림 2〕 채점자 수 증가에 따른 일반화가능도 계수, 의존도 계수

반면, 채점자를 고정한 상태에서 문항 수를 20개에서 25, 30, 35개로 늘리면 일반화가능도 계수 및 의존도 계수는 크게 높아진다. 예를 들어 국어 과목의 경우 채점자가 3명일 때, 문항 수를 20개에서 35개로 늘리면 일반화가능도 계수 및 의존도 계수는 각각 0.708과 0.620에서 0.809와 0.740으로 크게 높아지는 것에서 계수 값을 높이려면 문항 수를 늘이면 된다는 것을 〔그림 3〕에서 확인할 수 있다.



〔그림 3〕 문항 수 증가에 따른 일반화가능도 계수, 의존도 계수(채점자 3명)

나. 전국지역 국어 과목 분석

1) G연구 결과

〈표 7〉은 전국 및 16개 시·도 지역의 국어 과목을 대상으로 각각 $p \times i \times r$ 설계를 적용하여 분산 성분을 추정한 후 그 상대적 크기를 비교하여 설명하고 있다. 전국 시·도 지역 국어 과목의 분산 성분 추정치를 평균한 값을 보면, 전체 분산에 대해 전집점수 분산은 6.8%로 나타났고, 채점자 분산과 피험자와 채점자 간의 상호작용 분산은 각각 0.0%로 채점자에 따른 오차분산은 거의 존재하지 않음을 알 수 있다. 이는 채점자 그룹이 비교적 동질적으로 채점자가 채점 점수에 주는 변동 요인은 적다는 것을 의미한다. 문항 분산은 29.8%로 비교적 크게 나타났고, 반면 채점자와 문항 간의 상호작용 분산은 1.8%로 비교적 적게 나타났다. 피험자와 문항 간의 상호작용 분산은 54.7%로 대부분의 오차요인을 차지하고 있었고, 잔차 분산은 6.8%에 그쳐 상대적으로 적었다. 다만, 인천 지역의 경우 피험자와 문항 간의 상호작용 분산이 다른 지역에 비하여 적어 17.7%였고, 채점자와 문항 간의 상호작용 분산 및 잔차 분산은 각각 23.0%와 42.3%로 다른 지역에 비하여 큰 것으로 나타났다.

〈표 7〉 16개 시·도 지역 국어 과목 $p \times i \times r$ 설계에 대한

분산 성분 추정치 및 비율*

효과 / 지역	p	r	i	$p \times r$	$p \times i$	$r \times i$	$p \times i \times r$
서울	0.063 (7.3)	0.000 (0.0)	0.253 (29.2)	0.000 (0.0)	0.497 (57.4)	0.002 (0.2)	0.051 (5.9)
부산	0.072 (7.7)	0.000 (0.0)	0.297 (31.9)	0.000 (0.0)	0.504 (54.1)	0.003 (0.3)	0.056 (6.0)
대구	0.061 (6.8)	0.000 (0.0)	0.249 (27.9)	0.000 (0.0)	0.524 (58.7)	0.002 (0.2)	0.057 (6.4)
인천	0.068 (6.9)	0.000 (0.0)	0.100 (10.2)	0.000 (0.0)	0.174 (17.7)	0.226 (23.0)	0.415 (42.2)
광주	0.052 (5.6)	0.000 (0.0)	0.343 (36.8)	0.000 (0.0)	0.514 (55.2)	0.002 (0.2)	0.020 (2.1)
대전	0.058 (7.0)	0.000 (0.0)	0.231 (27.9)	0.000 (0.0)	0.497 (60.0)	0.002 (0.2)	0.041 (4.9)
울산	0.041 (4.5)	0.000 (0.0)	0.332 (36.2)	0.000 (0.0)	0.523 (57.10)	0.002 (0.2)	0.018 (2.0)
경기	0.056 (6.1)	0.000 (0.0)	0.297 (32.2)	0.000 (0.0)	0.518 (56.2)	0.006 (0.7)	0.045 (4.9)

강원	0.065 (7.6)	0.000 (0.0)	0.260 (30.5)	0.000 (0.0)	0.501 (58.8)	0.002 (0.2)	0.024 (2.8)
충북	0.054 (6.2)	0.000 (0.0)	0.263 (30.3)	0.000 (0.0)	0.509 (58.6)	0.002 (0.2)	0.041 (4.7)
충남	0.103 (11.8)	0.001 (0.1)	0.225 (25.8)	0.000 (0.0)	0.490 (56.1)	0.009 (1.0)	0.045 (5.2)
전북	0.072 (8.2)	0.000 (0.0)	0.273 (31.2)	0.000 (0.0)	0.491 (56.2)	0.002 (0.2)	0.036 (4.1)
전남	0.047 (5.3)	0.000 (0.0)	0.265 (30.0)	0.000 (0.0)	0.520 (58.9)	0.004 (0.5)	0.047 (5.3)
경북	0.055 (6.0)	0.000 (0.0)	0.315 (34.6)	0.000 (0.0)	0.496 (54.5)	0.004 (0.4)	0.040 (4.4)
경남	0.058 (6.5)	0.000 (0.0)	0.290 (32.6)	0.000 (0.0)	0.497 (55.8)	0.004 (0.4)	0.041 (4.6)
제주	0.046 (5.3)	0.000 (0.0)	0.258 (29.8)	0.000 (0.0)	0.521 (60.2)	0.007 (0.8)	0.033 (3.8)
전국	0.061 (6.8)	0.000 (0.0)	0.266 (29.8)	0.000 (0.0)	0.486 (54.7)	0.017 (1.8)	0.063 (6.8)

*( ) 안의 수치는 각 분산 성분 추정치가 전체검사점수 분산에서 차지하는 비율임.

2) D연구 결과

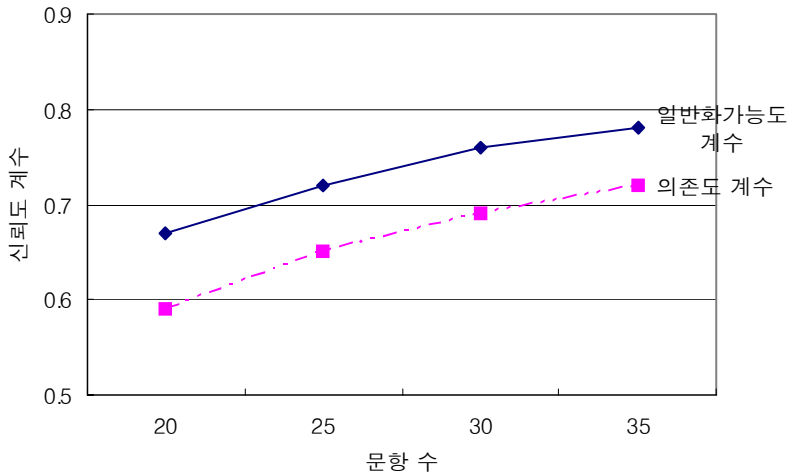
채점자와 문항 수를 조절하며 일반화가능도 계수에 미치는 영향을 알아보기 위해 다양한 D연구가 수행되었다. 채점자 수는 2, 3, 4, 5명일 경우로 조정하였고, 문항 수는 20, 25, 30, 35개로 변화시키며 분석하였다. 각각의 D연구 결과는 〈표 8〉에 제시되어 있다. D연구를 위한 분산 성분 추정치는 16개 시·도를 통합한 전국자료를 사용하였다. 각 시·도별로 분석을 위한 연구가 가능하고, 실제 이 연구에서 시행해 본 결과 전국자료를 사용한 결과와 유사하였다.

실제 A국가고사의 채점 조건을 반영하는 채점자를 3명으로, 문항 수를 30개로 하였을 때의 일반화가능도 계수는 0.76이고, 의존도 계수는 0.69로 서울지역 5개 과목의 계수보다 낮게 나타났다. 또한 서울지역의 분석과 마찬가지로 채점자의 수를 2명에서 3, 4, 5명까지 늘리더라도 일반화가능도 계수 및 의존도 계수는 거의 변하지 않았다. 반면, 〔그림 4〕에서 볼 수 있듯이 채점자를 고정한 상태에서 문항 수를 20개에서 35개로 늘리면 일반화가능도 계수 및 의존도 계수는 크게 높아졌다. 채점자가 3명일 때, 문항 수를 20개에서 35개로 늘리면 일반화가능도 계수 및 의존도 계수는 각각 0.67과 0.59에서 0.78과 0.72로 높아지는 것을 확인할 수 있다.

〈표 8〉 국어 과목 16개 시·도 종합 자료의 채점자 수와 문항 수 변화에 따른 신뢰도 계수

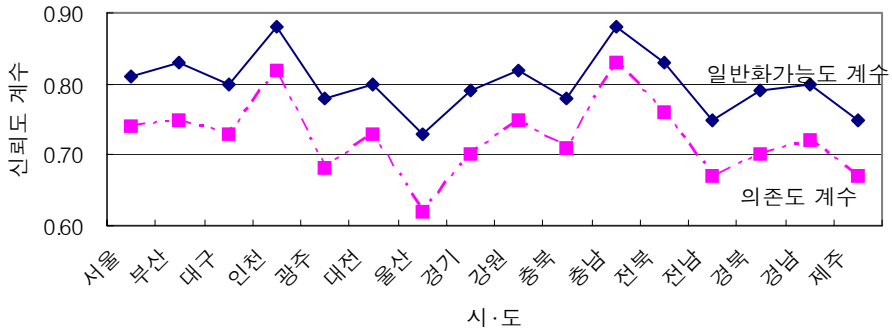
채점자 수	문항 수	전집점수 분산	관찰점수 분산	상대오차 분산	절대오차 분산	일반화가능도 계수	의존도 계수
2	20	0.05930	0.08832	0.02901	0.04110	0.67150	0.59064
	35	0.05930	0.07589	0.01659	0.02350	0.78141	0.71621
3	20	0.05930	0.08794	0.02863	0.04072	0.67440	0.59291
	35	0.05930	0.07567	0.01637	0.02327	0.78369	0.71815
4	20	0.05930	0.08775	0.02844	0.04053	0.67586	0.59406
	35	0.05930	0.07556	0.01626	0.02316	0.78484	0.71913
5	20	0.05930	0.08763	0.02833	0.04041	0.67674	0.59475
	35	0.05930	0.07550	0.01619	0.02310	0.78553	0.71972

※ 문항 수 변화는 20문항과 35문항으로만 표시함.



〔그림 4〕 국어 과목 16개 시·도 종합 자료의 문항 수에 따른 신뢰도 계수 비교(채점자 3명)

국어 과목의 시·도별 일반화가능도 계수 및 의존도 계수를 비교하여 보면, 서울지역 5개 과목의 경우와 마찬가지로 전체적으로 일반화가능도 계수 및 의존도 계수는 높은 수준임을 알 수 있다. 일반화가능도 계수가 0.80 이상인 시·도는 가장 수치가 높은 인천과 충남을 비롯하여 서울, 부산, 강원, 전북의 6개 시·도이며, 나머지 시·도의 경우도 모두 0.70 이상의 수치를 기록하였다. 이 정도 수준의 신뢰도 계수는 채점자를 포함하고 있는 측정 상황에서는 상당히 높은 수준의 신뢰도로 해석된다. 시·도별 국어 과목에 대한 일반화가능도 계수와 의존도 계수는 〔그림 5〕와 같다.



〔그림 5〕 국어 과목 16개 시·도별 신뢰도 계수 비교

Ⅳ. 논의

이 연구는 현행 A국가고사의 전공(주관식) 시험의 채점자가 신뢰도에 미치는 영향을 알아보기 위하여 일반화가능도 분석을 실시하였다. 검사의 일반적 특성을 알아보기 위한 문항분석 결과, 난이도 수준은 0.2에서 0.6 사이에 주로 분포하고 있었고, 이는 선발시험이라는 것을 감안할 때 좀 어려운 문항 구성으로 보인다. 문항변별도 분석에서는 비교적 양호한 분포를 나타냈으나 국어의 경우 0.2 미만의 변별도를 보인 문항이 8개로 상대적으로 많았다.

A국가고사는 그 본래의 목적을 달성하기 위해 같은 지역 내, 동일 교과의 지원자들은 동일한 채점자들에 의해 채점되기 때문에 채점자 효과가 선발에 미치는 영향을 적절히 통제하고 있는 것으로 보인다. 또한 복수 채점을 하고 있어 채점자 내 신뢰도 문제로 인한 선발의 불공정도 통제하려는 의도가 있음을 긍정적으로 평가할 수 있다. 이러한 채점 방식은 선발시험에서 채점자의 영향력을 가능한 한 균등하게 분배시켜 선발의 공정성에 기여할 것으로 보인다.

서울지역 5개 과목과 16개 시·도 국어 과목의 일반화가능도 분석에서 공통적으로 나타난 현상은 채점자 효과나 채점자와 관련된 효과, 예를 들어 채점자와 문항의 상호작용 효과, 채점자와 피험자의 상호작용 효과 등이 거의 0에 가깝다. 이러한 결과는 채점자 간에 상당한 수준의 일관성을 확보하고 있다는 증거가 된다. 이러한 일반화가능도 연구 결과는 채점자 간 상관계수 산출을 통해 살펴본 채점자 간 신뢰도 분석에서도 확인되었다.

채점자 간 일관성 정도를 실증적으로 알아보기 위한 상관분석 결과, 가장 낮은 상관을 보

인 국어 과목(r=.98)을 포함한 다른 모든 과목에서 상관계수가 거의 1에 가까운 상관을 보여 상당히 높은 수준의 채점자 간 일관성을 확인할 수 있었다. 이러한 결과는 채점을 시행하기 전 가채점 등을 활용한 채점자 훈련을 통해 채점자 간 일관성이 높아진 때문인 것으로 해석할 수도 있지만, 너무 높게 나타난 채점자 간 일관성이 채점조 운영의 문제에서 기인한 것인지에 대한 체계적인 검토가 필요할 것이다.

이 연구를 통해 극히 작게 나타난 채점자 효과는 일반적으로 채점자를 포함하고 있는 측정 상황에서는 발견되기 힘든 것이다. 즉 채점자 효과 분산은 작더라도 일반적으로 채점자와 문항의 상호작용이나 채점자와 피험자의 상호작용은 일정 수준의 분산이 발견되는 것이 일반적이다. 이 연구의 결과만으로는 거의 0에 가까운 채점자 효과가 어떤 이유에서 나타났는지를 결론 내리기 어렵다. 이러한 결과가 잘 정선된 채점 기준표와 구조화된 채점자 훈련의 효과 때문이라면 현재의 A국가고사에서 사용하고 있는 채점 기준표 작성 원칙이나 채점자 훈련 프로그램을 유지할 필요가 있다고 제안할 수 있을 것이다. 그러나 0에 가까운 채점자 효과가 채점조 운영 또는 자질의 문제에 기인하는 것이라면, 더 적합한 채점 절차를 구성할 수 있는 다양한 채점 디자인을 적용한 후속 연구가 이루어져야 할 것이다.

또한 이 연구에서 사용한 자료의 출처가 되는 국가고사를 밝히지 못하고 연구를 진행한 점은 분명히 적용 연구의 가치를 떨어뜨리는, 실제적 유의미성을 의심받는 것으로 이 연구의 제한점임에는 틀림이 없다.

참 고 문 헌

강애남, 이규민 (2006). 동료평가를 활용한 수행평가 결과의 일반화가능도 분석. **2006년도 교육평가학회 춘계학술대회 발표논문**.

김성숙 (1995). 논술문항 채점의 변동요인 분석과 일반화가능도 계수의 최적화 조건. **교육평가연구, 8**(1), 35-57.

김성숙, 김양분 (2001). **일반화가능도 이론**. 서울: 교육과학사.

성태제 (2002). **현대교육평가**. 서울: 학지사.

이규민 (2002). 의학교육의 대안적 평가: 심층면접. **제12차 의학교육합동학술대회 발표 논문**.

이규민 (2002b). A comparison of five scoring modalities in terms of reliability and efficiency. **교육평가연구, 15**(1), 227-245.

이규민 (2003). 단위검사 개념의 적용: 일반화가능도 이론을 중심으로. **교육평가연구, 16**(1), 53-70.

이규민 (2004). A generalizability theory approach to investigating item context effects. **교육평가연구, 17**(1), 219-237.

이영식, 신상근 (2004). 다변량 일반화가능도 이론에 의한 말하기 시험의 타당도와 신뢰도에 관한 연구. *Foreign Languages Eduaction, 11*(2). 6-14.

이종성 (1988). **일반화가능도이론**. 서울: 연세대학교출판부.

한국교육평가학회 (2004). **교육평가 용어사전**. 서울: 학지사.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: American College Testing.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Cronbach, L. J., Glesser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Gao, X., Brennan, R. L., & Shavelson, R. J. (1994). *Estimating generalizability of matrix-sampled science performance assessment*. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, Louisiana.

Gao, X., & Colton, D. A. (1996). Evaluating measurement precision of performance assessment with multiple forms, raters, and tasks. In M. J. Kolen (Chair), *Technical issues involving reliability and performance assessments*. Symposium conducted at the Annual Meeing of American Educational Research Association. New York City, New York.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*(4), 323-432.

Lee, G. (2002). The influence of several factors on reliability for complex reading comprehension tests. *Journal of Educational Measurement, 39*, 149-164.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice, 19*, 9-15.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*, 237-255.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). on the stability of perfromance assessment. *Journal of Educational Measurement, 30*, 41-53.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

Shavelson, R. J., Gao, X., & Baxter, G. P. (1996). On the content validity of performance assessments: Centrality of domain specifications. In M. Birenebaum & F. J. R. C. Cochy (Ed.), *Alternatives in assessments of achievements, learning processes and prior knowledge* (pp. 131-141). Boston, MA: Kluwer Academic Publishers.

• *논문 접수 : 2008년 2월 11일 / 수정본 접수 : 2008년 4월 4일 / 게재 승인 : 2008년 4월 15일*

ABSTRACT

A Generalizability Theory Approach to Investigating Rater Effects with 'A' National Exam Composed of Constructed-Response Items

Chang-Hwan Kim(Doctoral Student, Korea University)
Gue-Min Lee(Professor, Yonsei University)

Generalizability theory offers an extensive conceptual framework and a powerful tool for explaining numerous measurement issues and multiple sources of errors that classical test theory cannot clearly differentiate. It has another strength in quantifying the influence of rater effects on observed scores, especially tests composed of constructed-response items. However, there has been relatively few studies applying G-theory approaches in practice because most national tests including certificate and licensure exams have been administered in the form of multiple-choice items in Korea. Fortunately, we can find and use 'A' national exam including constructed-response items that require the involvement of raters. A generalizability theory approach was implemented to investigate rater effects on test scores and to evaluate the generalizability and dependability of those scores. The results of this study suggest that raters or rater-related effects (e.g., interaction effects of raters and items, interaction effects of persons and raters) are nearly zero. If these findings were due to well-established scoring rubric and well-structured rater training effects, it would be recommended to maintain current scoring rubric and rater training procedures. However, if these finding were resulted from the way of coordinating rater groups or the participated raters' ability, it could be recommended to study other scoring methods to ensure more proper and objective test results.

Key Words : generalizability theory, multiple sources of errors, constructed-response items, reliability involving raters