교육과정평가연구
The Journal of Curriculum & Evaluation
2004, Vol. 7, No. 1, pp. 255~273

일반화가능도 이론을 이용한 과학 수행평가의 오차원 분석 및 신뢰도 추정

이 기 영(한성과학고등학교 교사)
안 희 수(서 울 대 학 교 교 수)

≪ 요 약 ≫

본 연구에서는 일반화가능도 이론을 이용하여 과학 수행평가의 오차원을 분석하고, 이를 토대로 일반화가능도(신뢰도)를 추정하였다. 서울시 소재 고등학교 1학년 2개 반 90명의 과학 수행평가 자료를 대상으로 두 개의 일반화가능도 연구(G 연구)를 학기별로 실시하였다. 그 결과 서술형 문항(i)과 채점자(r)를 국면으로 하는 $p\times(i:r)$ 설계에서는 문항보다 채점자 관련 분산성분이 오차원에 더 크게 기여하는 것으로 추정되었으며, 반에 따라 채점자 관련 분산성분에 많은 차이가 있었다. 또 1학기에 비해 2학기의 채점자 관련 분산성분이 많이 감소하였는데, 이것은 학기를 거치면서 채점자들의 채점 능력이 향상된 훈련의 효과로 판단되었다. 수행과제(t)와 채점자(r)를 국면으로 하는 $p\times(t:r)$ 설계에서는 채점자보다 수행과제 관련 분산성분이 더 크게 추정되었으며, 반에 따라 채점자 관련 분산성분에 많은 차이가 있었다. 두 개의 G 연구 결과, 피험자가 어느 채점자군에 속하느냐에 따라, 또 어떤 과제를 수행하느냐에 따라 측정점수가 다를 수 있는 것으로 나타났다. G 연구 설계와 동일하게 실시한 결정 연구(D 연구) 결과, 대부분의 경우에서 일반화가능도 계수가 적정 수준인 0.8에 미치지 못하였으며, 적정 수준의 일반화가능도 계수를 얻기 위해서는 더 많은 수의 채점자와 문항 그리고 수행과제가 필요한 것으로 분석되었다. 또한 근본적인 처방으로 과학 수행평가의 일반화가능도를 높이기 위해서는 채점자간의 차이를 줄이기 위한 심도 있는 교사 훈련이 필요하며, 서술형 문항과 수행 과제간의 차이를 줄이기 위한 노력이 있어야 할 것으로 판단되었다.

주요어 : 과학 수행 평가, 신뢰도, 일반화가능도 이론, 채점자, 서술형 문항, 수행 과제

Ⅰ. 서론

수행평가는 학습자의 반응을 평가하기 위해서 관찰이나 교사의 판단에 크게 의존할 수밖

에 없고, 또 표준화가 불가능하기 때문에 신뢰도가 높지 못하다는 문제를 안고 있다(Ruiz-Primo & Shavelson, 1996). 학교 현장에 수행평가가 도입된 이래로 평가의 신뢰도는 끊임없이 제기되고 있는 문제 중의 하나이다. 특히 과학 교과에서의 수행평가는 그 평가 도구 및 수행 과제가 다양하기 때문에 평가의 신뢰도는 매우 중요한 문제이다. 현재 우리나라에서 실시되고 있는 과학 수행평가는 학생들의 성취도에 30~50% 정도 반영되는 것으로 보고되고 있으나, 학교 현장 사정이나 제반 여건의 미흡으로 인해 평가의 신뢰도가 의문시된다. 그러므로 학교 과학 수행평가가 과연 어느 정도 신뢰롭게 실시되고 있는지, 또 어느 정도 일반화가 가능한지, 그리고 수행평가가 일반화가능성 있는 신뢰로운 평가 도구가 되기 위해서는 어떤 요건을 갖추어야 하는지에 대한 연구가 있어야 할 것이다. 그런데 기존의 고전검사이론(classical test theory)을 통해 산출되는 신뢰도는 측정 도구의 신뢰도 즉, 측정 결과의 일관성(consistency)에 대한 추정 방법에 집중되어 왔기 때문에 관찰대상과 관찰자, 시기, 환경 및 상황 등의 오차원(sources of error)을 복합적으로 고려하지 못한 약점을 가지고 있었다. 다시 말해, 단순히 관찰대상의 검사 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 다중오차분산원(multiple sources of error variance)의 크기와 이들 간의 상호작용 효과를 동시에 추정할 수 있게 하기 위해 등장한 것이 바로 일반화가능도 이론(generalizability theory)이다(Cronbach et al., 1972).

고전검사이론에서 진점수(true score)는 여러 시기에 여러 번에 걸쳐 얻어진 불변의 개인 점수로, 무한히 반복 측정한 관찰점수의 평균 점수로 산출되며, 신뢰도 계수는 측정이 어느 정도 관찰대상의 진점수를 반영하는가를 의미한다(허경철, 1986). 그러므로 고전검사이론은 진점수를 고정시킴으로써 측정상황에서 발생할 수 있는 오차요인을 구체적으로 설명하지 못한다는 약점을 가진다. 반면, 일반화가능도 이론은 진점수 대신 전집점수(universe score)를, 신뢰도 계수 대신 일반화가능도 계수(G 계수; generalizability coefficient) 개념을 도입한다. 일반화가능도 이론은 관찰된 점수 하나하나가 전집(universe)을 대표하는 일부분이라고 전제하고, 전집을 구성하는 국면(facet)을 바탕으로 표집된 점수가 얼마나 일반화될 수 있는가를 추정하는 것으로, 전집은 오차원을 이루는 하나 이상의 국면으로 규정지어지며, 각 국면과 관련된 독립적인 분산성분(variance component)을 추정하여 상대적 크기를 파악함으로써 그 영향력을 비교한다(Shavelson et al., 1989). 일반화가능도 이론은 분산분석(ANOVA) 체계를 적용하여 측정상황에서 발생할 수 있는 다중오차요인을 동시에 분석하고, 측정점수에 대한 오차요인의 상대적 영향력을 산출하여 일반화가능도 계수와 함께 의사결정자에게 안정적인 점수를 얻기 위한 측정조건을 제시함으로써 신뢰도 추정과정을 한 단계 향상시킨 것이다(김성숙・김양분, 2001).

일반화가능도 이론은 크게 G 연구(generalizability study, 일반화가능도 연구)와 D 연구(decision study, 의사결정연구 또는 결정연구)로 나뉘어진다. G 연구는 측정 표본을 측정 전집(universe)에 일반화하고자 할 때, 허용가능한 관찰전집과 분산성분 추정치를 산출하는 과정으로 적절한 일반화가능도 계수를 얻기 위한 D 연구를 도와준다. D 연구는 연구자가 일반화하고자 하는 전집을 규정하고 측정에서 변동에 영향을 주는 관찰 조건을 조정하거나 결정하는 과정이다(Brennan & Webb, 1991; Brennan, 1992). G 연구설계의 목적은 허용가능한 관찰전집과 관련된 분산성분의 추정치를 구하는 것이다. G 연구에서 추정된 분산성분의 추정치는 다양한 D연구에서 측정대상에 대한 실질적인 의사결정을 위한 정보를 제공하거나 수행가능한 측정절차를 설계하는데 사용될 수 있다. 일반적으로 D연구는 잘 설계된 측정절차를 통해 의사결정을 하기 위한 분산성분의 추정, 사용, 해석을 강조한다(Brennan, 1983; Crocker, 1986).

일반화가능도 이론은 고전검사이론의 신뢰도 추정 방법에 비해 측정의 목적에 따라 진점수가 달라지며, 그에 따라 여러 개의 신뢰도가 존재한다. 또한 다중 오차 요인과 그 영향력을 규명할 수 있으며, 일반화가능도 계수와 함께 준거지향검사를 위한 의존도 계수도 동시에 추정할 수 있어, 논술형 문항의 채점, 예체능 실기 검사, 수업 중 교사 또는 학생 행동의 관찰 등에 광범위하게 활용되고 있다.

Kim(1989)은 일반화가능도 이론을 이용하여 교사행위 관찰에 있어서 오차원을 분석하였다. 관찰대상인 교사행위 중 선택된 5가지 영역을 측정함에 있어 관찰자(r)와 관찰 사례(o) 및 상호작용으로 인한 오차분산의 상대적 크기로 각 오차원의 영향력을 비교하였고, 적정 수준의 일반화가능도 계수 산출을 위해 필요한 관찰자 수와 사례의 수를 조합, 추정하여 5가지 행위영역을 비교하였다. 또 김성숙(1995)는 논술 문항 채점의 변동요인을 분석하여 일반화가능도 계수의 최적화 조건을 조사하였다. Brennan & Johnson(1995)은 1989-1990년의 CAP (California Assessment Program) 자료를 대상으로 채점자와 수행과제 국면이 오차 분산에 어떻게 기여하는지 조사하여, 채점자보다는 수행과제와 관련된 분산성분이 오차에 더 크게 기여한다는 결과를 얻었다. Ruiz-Primo(1993) 등은 수행평가 점수의 안정성(stability)을 검사하기 위해 3가지 탐구과제를 5개월 간격으로 2번에 걸쳐 실시하여, 학생들의 점수가 사례(occasion)에 따라 달라짐을 확인하였다. McClure, Sonak & Suen(1999)은 채점 방법이 개념도(concept map) 평가의 신뢰도에 미치는 영향을 조사하는데 일반화가능도 이론을 적용하였다. 그 결과, 채점 방법에 따라 신뢰도가 영향을 받는데, 이것은 각 채점 방법마다 채점자의 작동기억(working memory)에 부과되는 인지부하(cognitive load)가 다르기 때문인 것으로 분석되며, 채점자가 이러한 복잡성을 효과적으로 다루는 전략을 수립하지 못한다면 일관성(consistency)이 떨어져 신뢰도를 감소시키는 결과를 초래하게 된다고 보고하였다. Swartz et al.(1999)은 총체적(holistic) 방법과 분석적(analytic) 방법으로 채점한 쓰기 점수(writing score)의

신뢰도를 추정하는데 일반화가능도 이론을 적용하여, 채점자수 증가에 따라 G 계수가 증가한다는 결과를 얻었다. 그러나 비용의 문제가 있으므로 채점자 훈련 개선 또는 훈련 시간 증가를 채점자수를 늘려야하는 문제의 대안으로 제시하면서 채점자로서 심도 있는 교사 훈련이 중요함을 시사하였다. Sugrue, Webb & Schlackman(2000)는 과학 수행평가 점수의 일반화가능도와 과학 평가 방식의 교환가능성(interchangeability)을 추정함에 있어 오차분산원으로서 사례(occasion)의 중요성을 조사하였다.

위 선행 연구들을 살펴보면 수행평가에 대한 연구는 많으나 학교 현장의 자료를 대상으로 한 연구가 없고 대부분이 연구를 목적으로 인위적으로 구성되거나 채점 과정에 연구자가 개입한 것들이었다. 그러나 본 연구는 실제 학교 현장에서 실시한 과학 수행평가 자료를 대상으로 오차원들을 분석하여 그 영향력을 비교하고, 이를 토대로 신뢰도(일반화가능도)를 추정함으로써 학교 과학 수행평가에 어떤 요인들이 어느 정도 영향을 주고 있으며, 또 그 점수는 어느 정도 일반화가 가능한지 알아보고자 하였다.

본 연구의 연구 문제는 다음과 같다.

첫째, 과학 수행평가에서 피험자 점수의 오차분산에 기여하는 채점자(rater), 문항(item), 수행과제(task) 국면에 의한 분산 성분 추정치의 크기는 어떻게 나타나는가?

둘째, 오차분산에 기여하는 국면의 수를 조절함으로써 일반화가능도 계수는 어떻게 향상되는가?

셋째, 과학 수행평가의 일반화가능도를 어떻게 높일 수 있는가?

Ⅱ. 연구 내용 및 방법

1. 연구 대상

본 연구는 서울시 양천구 소재 고등학교 1학년 2개 반 학생 90명의 과학 수행평가 점수를 대상으로 분석하였다. 연구의 특성상 연구 대상반이 동질 집단일 필요는 없으므로 무선 표집 되었으며, 반1과 반2를 채점한 2명씩의 교사가 서로 다르기 때문에 반1과 반2로 분리하여 1학기와 2학기 수행평가 점수를 각각 분석하였다. 한 학기의 수행평가는 지필평가인 서술형 문항 10개와 실기평가인 수행과제 8개로 구성된다.

2. 연구 설계

가. G 연구 설계

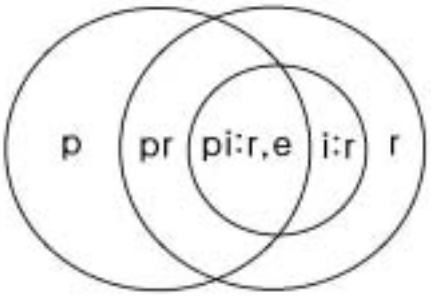
1) p×(i : r) **i=5. r=2**

이 설계는 1국면 교차, 1국면 내재 설계로, 모든 피험자(p)는 모든 서술형 수행평가 문항(i)을 치르고, 모든 채점자(r)가 모든 피험자를 채점하되 각 채점자는 문항을 나누어서 서로 다른 문항을 채점하는 설계이다. 본 연구에서는 10개의 서술형 문항을 2명의 채점자가 5문항씩 채점한 결과를 이용하였다. 채점 기준(scoring rubrics)은 각 문항당 0-3 scale을 사용하여 3, 2, 1, 0점으로 채점하였다.

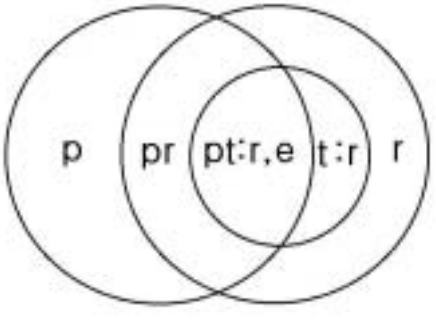
2) p×(t : r) **t=4, r=2**

이 설계 또한 1국면 교차, 1국면 내재 설계로, 모든 피험자(p)는 모든 수행과제(t)를 치르고, 모든 채점자(r)가 모든 피험자를 채점하되 각 채점자는 수행과제를 나누어서 서로 다른 수행과제를 채점하는 설계이다. 본 연구에서는 총 8개의 수행과제를 2명의 채점자가 4개씩 채점한 결과를 이용하였다. 채점 기준은 각 과제당 0-5 scale을 사용하여 5, 4, 3, 2, 1, 0점으로 채점하였다.

G 연구 설계를 벤 다이어그램으로 나타내면 [그림 Ⅱ-1]과 [그림 Ⅱ-2]와 같다. 그림에서 실선으로 구분되는 5개의 영역(section)은 G 연구를 통해 산출되는 5개의 분산성분을 나타낸다.



〔그림 Ⅱ-2〕 p×(i : r) 설계의 분산성분



〔그림 Ⅱ-2〕 p×(t : r) 설계의 분산성분

나. G 연구 분산성분

첫 번째 설계인 p×(i : r)의 경우 채점한 점수의 분산 성분은 아래 식으로 설명될 수 있다.

$$\sigma^2_{X_{pir}} = \sigma^2_p + \sigma^2_r + \sigma^2_{i:r} + \sigma^2_{pr} + \sigma^2_{pi:r,e}$$

σ^2_p : 피험자간의 차이

σ^2_r : 채점자간의 차이

$\sigma^2_{i:r}$: 채점자 내 문항간의 차이

σ^2_{pr} : 채점자에 따라 피험자를 다르게 채점한 정도

$\sigma^2_{pi:r,e}$: 잔차(residual) 또는 비체계적인(nonsystematic) 변량으로, 설명할 수 없는 변량

두 번째 설계인 p×(t : r)의 경우는 위 식에서 r대신에 t를 대입하면 된다.

다. D 연구 설계

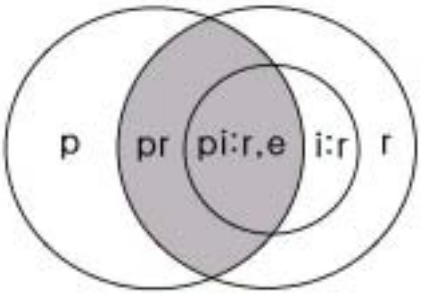
D 연구 설계는 G 연구 설계와 같은 p×(I :R)과 p×(T :R)로 하였으며, 채점자와 문항 그리고 수행과제 국면은 모두 무선효과(random effect)를 가정하였다. D 연구에서 오차분산은 상대적 결정(relative decision)에서 사용되는 상대오차분산(σ^2_δ)과 절대적 결정(absolute decision)에서 사용되는 절대오차분산(σ^2_Δ)으로 구분한다. 상대적 결정은 규준지향검사(norm-referenced test)처럼 개인의 점수를 상대적 위치로 해석하는 경우로, 이 때 상대 오차분산은 관찰점수와 전집점수 평균에 대한 상대적 차이를 의미하므로 측정대상을 포함하고 있는 상호분산성분만을 포함한다. 반면, 절대적 결정은 준거지향검사(criterion-referenced test)처럼 개인의 점수를 절대 기준에 의해 도달 여부를 파악하는 경우로, 이 때 절대오차분산은 측정대상의 관찰점수와 전집점수 차이에 대한 분산이므로 측정대상 분산을 제외한 모든 분산성분을 포함한다. 일반적으로 일반화가능도(generalizability) 계수의 산출은 상대오차분산을 이용하며, 절대오차분산은 준거지향검사의 의존도(dependability) 계수의 산출에 이용된다.

상대적 결정에서 일반화가능도 계수(ρ^2)와 절대적 결정에서 의존도 계수(Φ)는 아래 식과 같이 계산되며, 이 때 오차분산 σ^2_δ와 σ^2_Δ은 각각 [그림 Ⅱ-3]과 [그림 Ⅱ-4]의 벤다이어그램에서 어두운 부분에 해당된다. 일반적으로 절대오차분산은 상대오차분산보다 크기 때문에 의존도 계수는 일반화가능도 계수보다 언제나 작게 계산된다.

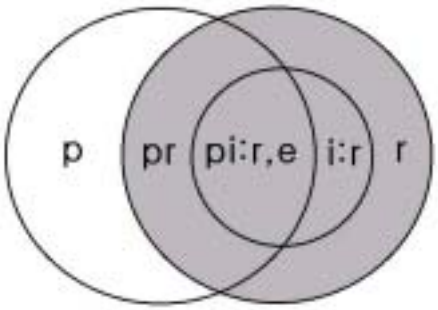
$$\rho^2 = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\delta} \quad \left(\sigma^2_\delta = \sigma^2_{pR} + \sigma^2_{pI:R,e} = \frac{\sigma^2_{pr}}{N_r} + \frac{\sigma^2_{pi:r,e}}{N_i N_r} \right)$$

$$\Phi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\Delta} \left(\sigma^2_\Delta = \sigma^2_R + \sigma^2_{I:R} + \sigma^2_{pR} + \sigma^2_{pI:R,e} \right.$$

$$\left. = \frac{\sigma^2_r}{N_r} + \frac{\sigma^2_{i:r}}{N_i N_r} + \frac{\sigma^2_{pr}}{N_r} + \frac{\sigma^2_{p\,i:r,e}}{N_i N_r} \right)$$



〔그림 II-4〕 상대적 결정의 오차분산



〔그림 II-4〕 절대적 결정의 오차분산

3. 연구 분석

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).

Ⅲ. 결과 및 논의

1. G 연구 결과 및 분석

가. p×(i : r) 설계의 분산 성분

반1과 반2의 각 분산성분 추정치와 그 비율을 학기별로 <표 Ⅲ-1>과 <표 Ⅲ-2>에 제시하였는데, 여기서는 전집분산(σ^2_p)과 각 오차분산성분의 크기를 상대적으로 비교함으로써 변동 요인의 영향력을 설명하고 있다. 반1의 경우 잔차성분을 제외하면, 1학기에는 채점자간 차이를 나타내는 오차분산(σ^2_r)이 30.7%로 가장 크게 나타났고, 2학기에는 피험자간의 차

이를 나타내는 전집분산(σ^2_p)이 20.2%로 가장 크게 나타났으며, 오차분산 중에는 채점자 내 문항간의 차이를 나타내는 분산($\sigma^2_{i:r}$)이 15.3%로 크게 나타났다. 1학기에 30.7%로 가장 크게 나타났던 σ^2_r은 2학기에는 7.8%로 감소하였다. 반2의 경우는 잔차성분을 제외하면 1, 2학기 모두 전집분산이 가장 커 반1의 경우보다 일반화가능성이 높은 것으로 나타났다. 특히 2학기에는 채점자 관련 분산성분인 σ^2_r과 σ^2_{pr}이 현격하게 감소하였다. σ^2_r은 1학기에는 11.4%였는데 2학기에는 0.0%로 크게 감소하였으며, σ^2_{pr}은 6.6%에서 1.0%로 감소하였다. 1학기에 비해 2학기에 전집분산이 증가되고, 오차분산이 감소함으로써 일반화가능성이 향상되었음을 알 수 있다. 반1과 반2 모두 1학기보다 2학기에 채점자간 차이인 σ^2_r가 감소한 것으로 나타났는데, 이것은 크게 두 가지로 해석이 가능하다. 하나는 학기를 거치는 동안 채점자간의 차이가 줄어든 훈련의 효과로 해석할 수 있고, 다른 하나는 2학기 문항들의 채점 기준이 좀 더 쉽고 명확하게 제시된 경우를 포함하는 그 이외의 변인으로 해석할 수 있다. 연구 결과로만 보아서는 전자로만 해석하는 것이 타당할 것이다. 설명할 수 없는 변량인 잔차가 전체분산 중 가장 크다는 것은 채점자와 문항 국면 이외에 오차분산에 기여하는 또 다른 변동요인이 있음을 암시한다.

〈표 III-1〉 p×(i : r) 설계에서 산출된 반1의 학기별 분산 성분

변동 요인	분산 성분	추정치		%	
		1학기	2학기	1학기	2학기
피험자(p)	σ^2_p	0.259	0.375	12.1	20.2
채점자(r)	σ^2_r	0.658	0.145	30.7	7.8
i:r	$\sigma^2_{i:r}$	0.356	0.284	16.6	15.3
pr	σ^2_{pr}	0.149	0.203	7.0	10.9
잔차(pi:r,e)	$\sigma^2_{pi:r,e}$	0.719	0.850	33.6	45.8

〈표 III-2〉 p×(i : r) 설계에서 산출된 반2의 학기별 분산 성분

변동 요인	분산 성분	추정치		%	
		1학기	2학기	1학기	2학기
피험자(p)	σ^2_p	0.567	0.833	31.1	46.9
채점자(r)	σ^2_r	0.208	0.000	11.4	0.0
i:r	$\sigma^2_{i:r}$	0.186	0.208	10.2	11.7
pr	σ^2_{pr}	0.119	0.017	6.6	1.0
잔차(pi:r,e)	$\sigma^2_{pi:r,e}$	0.740	0.717	40.6	40.4

나. p×(t : r) 설계의 분산 성분

p×(t : r) 설계에서 산출된 반1과 반2의 각 분산성분 추정치와 그 비율은 <표 Ⅲ-3>과 <표 Ⅲ-4>와 같다. 반1과 반2 모두 오차분산 중 잔차성분이 전체분산의 61.6~76.1%로 매우 크게 나타났다. 이것은 채점자와 과제 국면 이외에 오차분산에 크게 기여하는 또 다른 변동 요인이 있음을 암시한다. 잔차성분을 제외하면, 반1과 반2 모두 오차분산 중 채점자 내 과제간의 차이를 나타내는 $\sigma^2_{t:r}$이 11.4~21.1%로 가장 크게 나타났는데, 이것은 과제간에 상당한 차이가 있음을 의미한다. 그러나 반1의 2학기는 예외로 0.0%로 나타난 채점자 내 과제간의 차이가 없는 것으로 나타났다. 반1에서는 채점자와 관련된 오차분산성분이 1학기보다 2학기에 증가되는 경향을 보이는데 반해, 반2의 경우는 채점자와 관련된 오차분산성분들이 모두 0.0%로 나타났는데, 이것은 반1과 반2의 채점자간에 상당한 차이가 있으며, 반2가 반1보다 동질의 채점자로 비교적 일관성 있게 채점되었음을 의미한다. 한편, 반1과 반2 모두 피험자간의 차이를 나타내는 전집분산 σ^2_p 이 10.5~15.5%로 오차분산에 비해 매우 작게 나타났다.

〈표 Ⅲ-3〉 p×(t : r) 설계에서 산출된 반1의 학기별 분산 성분 추정치

변동 요인	분산 성분	추정치		%	
		1학기	2학기	1학기	2학기
피험자(p)	σ^2_p	0.049	0.114	10.5	15.5
채점자(r)	σ^2_r	0.000	0.133	0.0	18.1
t:r	$\sigma^2_{t:r}$	0.053	0.000	11.4	0.0
pr	σ^2_{pr}	0.009	0.035	1.9	4.8
잔차(pt:r,e)	$\sigma^2_{pt:r,e}$	0.357	0.453	76.1	61.6

〈표 Ⅲ-4〉 p×(t : r) 설계에서 산출된 반2의 학기별 분산 성분 추정치

변동 요인	분산 성분	추정치		%	
		1학기	2학기	1학기	2학기
피험자(p)	σ^2_p	0.328	0.189	14.4	11.9
채점자(r)	σ^2_r	0.000	0.000	0.0	0.0
t:r	$\sigma^2_{t:r}$	0.480	0.240	21.1	15.1
pr	σ^2_{pr}	0.000	0.000	0.0	0.0
잔차(pt:r,e)	$\sigma^2_{pt:r,e}$	1.470	1.162	64.5	73.1

2. D 연구 결과 및 분석

가. p×(I :R) 설계의 일반화가능도 계수

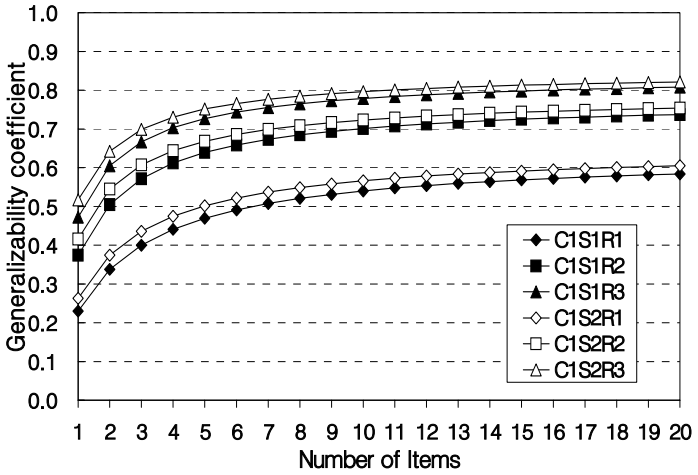
반1과 반2의 D 연구 실시 결과는 <표 Ⅲ-5>와 같다. 표는 G 연구 설계와 동일한 D 연구의 결과로 산출된 일반화가능도 계수를 정리한 것이다. 반1의 D 연구 결과, 각 국면의 기본조건(Nr=1, Ni=1)에 의한 1학기의 G 계수는 0.230으로 나타났으며, 2학기에는 G 계수가 약간 향상되는 결과를 보였다. 적정 수준의 G 계수에 도달하기 위해서 채점자와 문항 국면을 증가시킨 결과, 1학기의 경우 채점자 3명, 문항 16개 정도가, 2학기의 경우는 채점자 3명, 문항 11개 정도가 필요한 것으로 나타났다. 반2의 경우는 각 국면의 기본조건(Nr=1, Ni=1)에 의한 1학기의 G 계수가 0.398로 반1의 경우보다 높게 나타났으며, 2학기에는 G 계수가 많이 향상되는 결과를 보였다. 적정 수준의 G 계수에 도달하기 위해서는 1학기의 경우 채점자 2명, 문항 5개 정도가, 2학기는 채점자 2명, 문항 2개 정도가 필요한 것으로 나타났다.

〈표 Ⅲ-5〉 p×(I : R) 설계에서 산출된 반1과 반2의 학기별 일반화가능도 계수

국면의 수	G-계수(ρ^2)			
	반1(1학기)	반1(2학기)	반2(1학기)	반2(2학기)
Ni=1, Nr=1	0.230	0.262	0.398	0.532
G 연구의 Ni, Nr*	0.639	0.668	0.809	0.912
수준 0.80	Ni=16, Nr=3	Ni=11, Nr=3	Ni=5, Nr=2	Ni=2, Nr=2

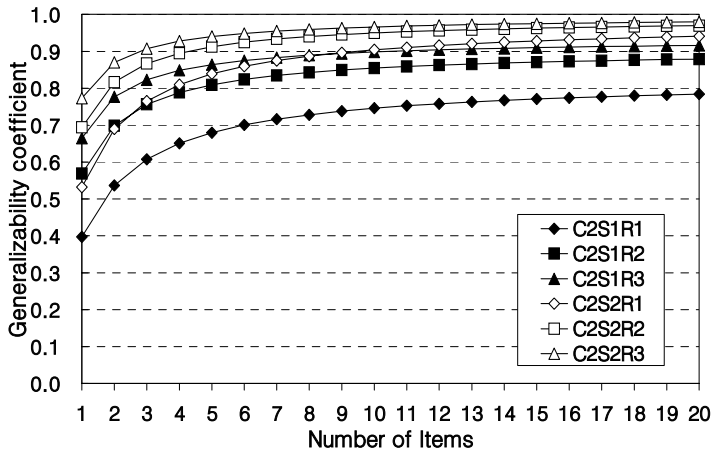
* G 연구에서 사용된 Ni=5, Nr=2 임

[그림 Ⅲ-1]은 반1의 학기별 문항과 채점자 수 증가에 따른 일반화가능도 계수 변화 추이를 그래프로 비교한 것으로 문항은 20개, 채점자는 3명인 경우까지 나타내었다. 여기서 C1S1R1은 반1(C1)에서 1학기(S1)에 채점자 1명(R1)인 경우를 의미한다. 그래프를 보면 1학기보다 2학기에 G 계수가 약간 증가되는 것을 볼 수 있다. 또한 채점자 수 증가에 따른 G 계수 변화 폭이 비해 문항 수 증가에 따른 그래프 기울기 변화가 매우 완만한 것으로 보아, 채점자 수의 증가가 문항 수의 증가보다 G 계수를 더 많이 향상시킨다는 것을 알 수 있다. 이것은 G 연구에서 채점자와 관련된 분산성분이 문항 관련 분산성분보다 더 크게 나왔기 때문이다.



〔그림 III-1〕 반1의 p×(I : R) 설계에서 문항과 채점자 국면 수 증가에 따른 G 계수의 변화

[그림 Ⅲ-2]는 반2의 학기별 문항과 채점자 수 증가에 따른 일반화가능도 계수 변화 추이를 그래프로 비교한 것이다. 1학기보다 2학기에 G 계수가 상당히 향상된 것을 볼 수 있으며, 채점자 수 증가에 따른 G 계수 변화 폭이 반1에서보다 작게 나타난다. 또 반1의 경우보다 문항 수의 증가가 G 계수를 더 많이 향상시킨다는 것을 알 수 있는데, 이것은 반1 보다 G 연구에서 문항과 관련된 분산성분이 더 크게 나왔기 때문이다.



〔그림 III-2〕 반2의 p×(I : R) 설계에서 문항과 채점자 국면 수 증가에 따른 G 계수의 변화

나. D 연구 p×(T :R)의 분산성분과 일반화가능도 계수

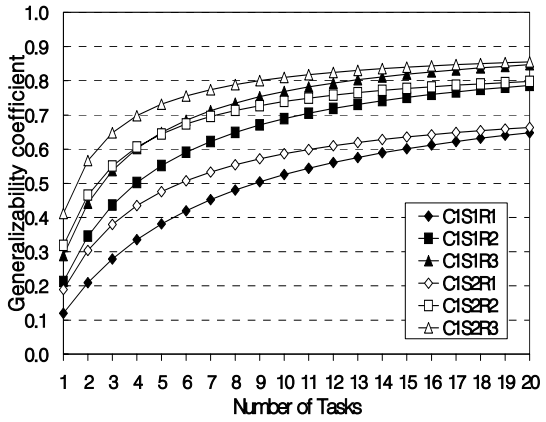
반1과 반2의 D 연구 결과는 <표 Ⅲ-6>과 같다. 반1의 D 연구 결과, 각 국면의 기본조건(Nr=1, Nt=1)에 의한 1학기의 G 계수는 0.119로 나타났으며, 2학기에는 0.189로 약간 향상되는 결과를 보였다. 적정 수준의 G 계수에 도달하기 위해서 채점자와 과제 국면을 증가시킨 결과, 1학기의 경우 채점자 3명에 과제 13개 정도가, 2학기의 경우는 채점자 3명에 과제 9개 정도가 필요한 것으로 나타났다. 반2의 경우는 각 국면의 기본 조건(Nr=1, Nt=1)에 의한 1학기의 G 계수가 0.182로 반1의 경우보다 약간 높게 나타났으나, 2학기에는 0.140으로 오히려 하락하는 결과를 보였다. 이것은 채점자 내 과제간 차이를 나타내는 분산성분이 감소하였음에도 불구하고, 전집분산은 감소하고 잔차분산성분이 증가하였기 때문으로 분석된다. 적정 수준의 G 계수에 도달하기 위해서 채점자와 과제 국면을 증가시킨 결과, 1학기의 경우 채점자 2명에 과제 9개 정도가 필요한 것으로 나타났고, 2학기는 채점자 2명에 과제 12개 정도가 필요한 것으로 나타났다.

〈표 Ⅲ-6〉 p×(T : R) 설계에서 산출된 반1과 반2의 학기별 일반화가능도 계수

국면의 수	G-계수(ρ^2)			
	반1(1학기)	반1(2학기)	반2(1학기)	반2(2학기)
Nt=1, Nr=1	0.119	0.189	0.182	0.140
G 연구의 Nt, Nr*	0.499	0.606	0.641	0.565
수준 0.80	Nt=13, Nr=3	Nt=9, Nr=3	Nt=9, Nr=2	Nt=12, Nr=2

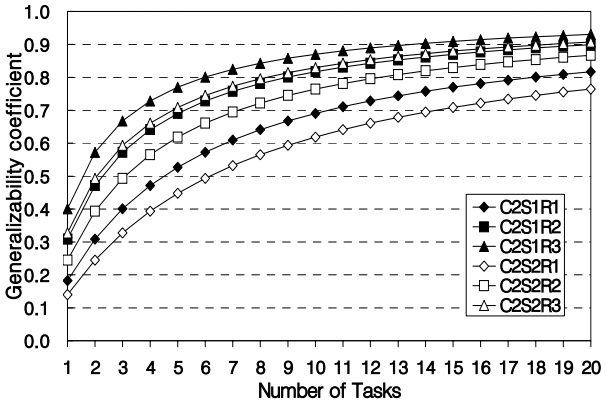
* G 연구에서 사용된 Nt=4, Nr=2 임

[그림 Ⅲ-3]은 반1의 학기별 수행과제와 채점자 수 증가에 따른 일반화가능도 계수 변화 추이를 그래프로 비교한 것이다. 전반적으로 1학기의 G 계수에 비해 2학기의 G 계수가 크게 나타나는 것을 확인할 수 있다. 또한 1학기와 2학기의 G 계수 변화 추이가 서로 다른 것을 볼 수 있다. 1학기는 과제 수 증가에 따른 그래프의 기울기가 2학기보다 큰 데 비해, 채점자 수 증가에 따른 G 계수 변화 폭은 1학기와 2학기가 거의 비슷하다. 이것은 1학기는 과제 수를 증가시키는 것이, 2학기는 채점자 수를 증가시키는 것이 G 계수를 향상시키는 데 보다 더 효과적임을 의미한다.



〔그림 III-3〕 반1의 p×(T : R) 설계에서 수행과제와 채점자 국면 수 증가에 따른 G 계수의 변화

[그림 Ⅲ-4]는 반2의 학기별 수행과제와 채점자 수 증가에 따른 일반화가능도 계수 변화 추이를 그래프로 비교한 것이다. 반1에서와는 달리 2학기에 G 계수가 오히려 감소하는 것을 볼 수 있다. 또한 1학기와 2학기 모두 채점자 수 증가에 따른 G 계수 변화 폭과 과제 수 증가에 따른 그래프 기울기 변화 양상이 거의 비슷하다. 1학기와 2학기 모두 과제 수 증가에 따른 그래프 기울기가 반1에서보다 큰 것으로 보아, 1학기와 2학기 모두 과제 수를 증가시키는 것이 채점자 수를 증가시키는 것보다 G 계수를 향상시키는 데 효과적이라는 것을 알 수 있다. 이것은 G 연구에서 채점자와 관련된 분산성분이 0.000이었기 때문인 것으로 판단된다.



〔그림 III-4〕 반2의 p×(T : R) 설계에서 수행과제와 채점자 국면 수 증가에 따른 G 계수의 변화

Ⅳ. 결론 및 제언

일반화가능도 이론은 단일오차원만을 고려하는 고전검사 이론을 확대하여 다중오차원을 고려하는 측정모형에 분산분석(ANOVA) 절차를 적용한 이론이라 할 수 있다(이종성, 1988). 이에 본 연구는 2가지 G 연구 설계로 고등학교 과학 수행평가의 다중오차원을 분석하고, 이를 토대로 신뢰도(일반화가능도)를 추정하는데 일반화가능도 이론을 이용하였다.

p×(i : r) 설계의 G 연구를 통해 2개 반 각각의 분산성분을 학기별로 추정한 결과, 반1과 반2의 분산성분의 크기와 비율이 상당한 차이를 보였으며, 1학기보다 2학기에 오차분산성분이 감소하는 것으로 나타났다. 대체적으로 문항보다 채점자 관련 분산성분이 오차원에 더 크게 기여하는 것으로 추정됨으로써 채점자간에 많은 차이가 있는 것으로 나타났다. 또 반에 따라 채점자 관련 분산성분에 많은 차이가 있었는데, 이것은 측정대상인 피험자들이 어느 채점자군에 속하느냐에 따라 측정점수가 다를 수 있다는 것을 의미한다. 1학기에 비해 2학기의 채점자 관련 분산성분이 많이 감소하였는데, 이것은 학기를 거치면서 채점자간의 차이가 줄어든 훈련의 효과로 판단할 수 있겠다.

p×(t : r) 설계의 G 연구에서는 대체적으로 채점자보다 수행과제 관련 분산성분이 오차원에 가장 크게 기여하는 것으로 추정되었으며, 반에 따라 채점자 관련 분산성분에는 많은 차이가 있었다. 이것은 수행과제간에 차이가 있어 같은 피험자라 할지라도 과제마다 측정점수에 차이가 있었음을 의미하며, 어느 채점자군에 속하느냐에 따라 측정점수가 다를 수 있다는 것을 의미한다. G 연구 결과를 종합해보면, 서술형 문항을 이용한 지필형 수행평가의 경우는 채점자간 차이가 많으며, 문항간에도 많은 차이가 있는 것으로 나타났다. 수행과제를 이용한 실기형 수행평가의 경우는 채점자보다는 수행과제간의 차이가 더 큰 것으로 나타났다. 그러나 설명할 수 없는 변량인 잔차성분이 매우 크게 나타남으로써 채점자와 수행과제 그리고 문항 관련 분산 이외에 다른 오차원의 분석이 필요한 것으로 나타났다.

p×(I : R) 설계의 경우 반1은 1, 2학기 모두 G 연구에 사용된 서술형 문항이나 채점자 수 정도로는 적정 수준의 G 계수에 도달하지 못하였으나, 반2는 1, 2학기 모두 G 연구에 사용된 서술형 문항이나 채점자 수 정도로 적정 수준의 G 계수에 도달할 수 있는 것으로 나타났다. 적정 수준의 G 계수인 0.8에 도달하기 위해서 채점자와 문항 국면을 증가시킨 결과, 반1의 경우는 채점자 수의 증가가 문항 수의 증가보다 G 계수 향상에 더 크게 기여한 반면, 반2의 경우는 문항 수의 증가가 채점자 수의 증가보다 G 계수 향상에 더 효과적인 것으로 나타났다. p×(T : R) 설계의 경우 G 연구에서 사용된 수행과제나 채점자 수 정도로는

적정 수준의 G 계수에 도달하지 못하였다. 적정 수준의 G 계수에 도달하기 위해서 채점자와 과제 국면을 증가시킨 결과, 과제 수의 증가가 채점자 수의 증가보다 G 계수 향상에 보다 효과적인 것으로 나타났다. D 연구에서는 채점자 수의 증가를 3명까지로 제한하였는데, 그 이유는 현실적인 비용 문제를 고려하여 실제 학교 현장에서 한 학생의 수행 문항이나 과제를 채점하는데 4명 이상의 채점자를 투입한다는 것은 비현실적이라고 판단하였기 때문이다.

채점자, 서술형문항, 수행과제를 과학 수행평가의 오차원으로 설계하여 분석한 결과, 일반화가능도를 높이기 위해서는 먼저 채점자간에 차이가 있으므로 채점자로서 심도있는 교사 훈련이 필요하며, 서술형문항간에도 차이가 있으므로 문항을 출제할 때 문항간의 내용이나 난이도 등이 너무 차이나지 않도록 보다 많은 노력을 기울여야 할 것으로 분석되었다. 또 수행과제를 구성하는데 있어서도 너무 이질적인 것이 되지 않도록 하는 것이 일반화가능도를 높일 수 있을 것으로 분석되었다. 그러나 설명할 수 없는 잔차성분이 너무 크게 나타나 또 다른 오차원의 분석이 요구되므로 평가기준(rubric), 평가영역(domain), 평가방법(method), 평가사례(occasion) 등을 오차원으로 포함하는 후속 연구가 있어야 할 것으로 보인다.

1개 고등학교의 자료를 대상으로 하였으므로 이것을 전체로 일반화시키는 데는 제한이 있지만, 본 연구의 결과 학교에서 실시하고 있는 과학 수행평가의 신뢰도가 낮은 것으로 분석되었다. 신뢰도가 낮다는 것은 평가하는 과정에서 일관성의 결여를 포함한 오차요인들이 개입되었다는 의미이므로 이 점수를 학생의 성취도에 그대로 반영하는 것은 수행평가의 본질을 훼손하는 중대한 문제가 될 수 있다. 지금까지의 과학 수행평가는 방법론적 측면만을 강조하며 타당한 성적 산출에 치중한 나머지 신뢰도 문제는 그리 심각하게 고려되지 못한 것이 사실이다. 그러므로 본 연구는 기존의 연구와는 달리 실제 학교 현장에서 실시한 과학 수행평가 자료를 대상으로 고전검사이론으로는 불가능한 다중오차원을 일반화가능도 이론을 이용하여 분석함으로써 각각의 오차원들이 과학 수행평가 점수에 얼마만큼 영향을 미치는지 알아내었으며, 과학 수행평가의 일반화가능도를 높이기 위한 방안을 결정하는데 필요한 구체적인 정보를 제공하였다는 점에서 그 의미가 있다고 하겠다.

참 고 문 헌

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

김성숙, 김양분(2001). **일반화가능도 이론**. 교육과학사.

이종성(1988). **일반화가능도 이론**. 연세대학교 출판부.

허경철(1986). 신뢰도에 관한 새로운 입장. **교육평가연구**, **1**(1), 183-194.

Brennan, R. L.(1983). *Elements of generalizability theory*. Iowa city, IA: American College Testing Program.

Brennan, R. L., & Webb, N. M.(1991). *Generalizability theory: A primer*. Thousand Oaks, California: Sage publications.

Brennan, R. L.(1992). Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4), 27-34.

Brennan, R. L., & Johnson, E. G.(1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*(4), 9-12.

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.

Crocker, L. M.(1986). Introduction to generalizability theory. *Introduction to classical and modern test theory*. Holt, Rinehart and Winston. Inc.

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.

Kim, S.(1989). *An analysis of source of variation in teacher behaviors using generalizability theory*. A dissertation presented to the graduate faculty of the university of virginia in candidacy for the degree of doctor of philosophy.

McClure, J. R., Sonak, B., & Suen, H. K.(1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching*, *36*(4), 475-492.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J.(1993). On the stability of performance assessments. *Journal of Educational Measurement*, *30*(1), 41-53.

Ruiz-Primo, M. A., & Shavelson, R. J.(1996). Rhetoric and reality in science performance assessment: an update. *Journal of Research in Science Teaching*, *33*, 1045-1063.

Shavelson, R. J., Webb, N. M., & Rowley, G. L.(1989). Generalizability theory. *American Psychologist*, *44*(6), 922-932.

Sugrue, B., & Webb, N., Schlackman, J.(2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education*, *13*(3), 277-301.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., & Wakely, M. B.(1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement*, *59*(3), 492-506.

• *논문접수 : 2004년 4월 15일 / 수정본 접수 : 2004년 5월 18일 / 게재 승인 : 2004년 6월 7일*

ABSTRACT

An Analysis of Sources of Error and an Estimation of Reliability in Science Performance Assessment using Generalizability Theory

Ki-Young Lee(Teacher, Hansung Science Highschool)
Hui-Soo An(Professor, Seoul National University)

In this study, we analyzed the sources of error underlying in science performance assessment by employing the generalizability theories, and its reliabilities estimated from this analyzed results, are presented and discussed. Science performance assessment data was obtained from two tenth-grade classes located in Seoul city, and two generalizability studies(G study) - $p\times(i:r)$ design and $p\times(t:r)$ design - is applied for both 1st and 2nd semesters. Where, facet i, t, and r indicate essay-type item, performance task and rater, respectively.

The results, in the $p\times(i:r)$ G studies, show that the variance components associated with rater(r) are found to be larger than those resulting from item(i). In addition, the estimated variance components showed a variety of differences from class to class. Variance components related to rater(r) were decreased apparently in the 2nd semester compared with the 1st semester. This result appears to be correlated with the increased training effect, mainly due to the enhancement of raters' scoring abilities in the 2nd semester.

On the other hand, the $p\times(t:r)$ study shows that the variance components originating from rater(t) were larger than those from item(r), and also variance components related to rater(r) show much differences from class to class.

These findings from two G studies indicate that person's score might differ according to rater group and performance task.

The results from Decision Study indicates that generalizability coefficients were turned out to be lower than the acceptable generalizability level(0.8).

We concluded that the population of statistical parameters such as number of rater, item and performance task, should be increased for approaching the acceptable level of generalizability coefficients, and basically, teacher training in rating and diminishing the difference in essay-type item and performance task are

also needed to improve the generalizability.

Key Words : science performance assessment, reliability, generalizability theory, rater, essay-type item, performance task.