

## 일반화가능도 이론을 이용한 집단논리적사고력검사(GALT)의 신뢰도 분석

류춘렬\* · 이용근

충북대학교 지구과학교육과, 361-763, 충북 청주시 흥덕구 성봉로 410

### An Analysis of the Reliability of Group Assessment of Logical Thinking (GALT) using Generalizability Theory

Chun-Ryol Ryu\* and Yong-Geun Lee

Department of Earth Science Education, Chungbuk National University, Chungbuk 361-763, Korea

**Abstract:** The purpose of this study lies in applying generalizability theory depending on the aim of the usage of GALT to analyze the sources of error of single-facet considering item and person only and to analyze the sources of error of multi-facet considering item, person and domain. The study was conducted with 1016 students of local elementary, middle, and high schools. The 21 items of a full version were answered for 40 minute and then the 12 items of short version were sampled to analyze reliability using generalizability theory. Both the full version and the short version of the items were analyzed using Cronbach's alpha for data analysis, and we applied generalizability theory and separate  $p \times i$  design and  $p \times (i:h)$  design, G study and D study were performed. Results of analysis are as follows: First, the result of D study after  $p \times I$  design both on the full version and the short version showed that in the case of the full version, the generalizability coefficient was 0.87 exceeding a normal level of 0.80, and the normal level of generalizability coefficient was achieved in 13 items as well. In case of short version, when 12 items were evaluated, generalizability coefficient was 0.77 not reaching the normal level, and the normal level was achieved in case of more than 15 items. Second, the result of D study after  $p \times (I:H)$  design on the short version showed that once one domain consists of 2 items in 6 domains, generalizability coefficient was 0.71 which is lower than the normal level of 0.80, the normal level was achieved in more than 5 item cases.

Keywords: GALT, generalizability theory, reliability

**요 약:** 본 연구의 목적은 집단논리적사고력검사의 이용 목적에 따라 일반화가능도이론을 적용하여 문항과 피험자만을 고려한 단일국면의 오차원, 그리고 문항과 피험자, 그리고 영역을 고려한 다국면의 오차원을 분석하는 데 있다. 연구는 지방 소재 초·중·고등학생 총 1016명을 대상으로 이루어졌으며, 21문항의 GALT 완본을 40분 동안 실시하고, 이 중 축소본에 해당하는 12문항을 별도로 추출하여 일반화가능도이론을 이용한 신뢰도 분석에 이용하였다. 자료의 분석을 위해 일반화가능도이론을 적용하여  $p \times i$ 설계와  $p \times (i:h)$ 설계로 나누어 G 연구와 D 연구를 실시하였다. 분석결과는 다음과 같다. 첫째, 완본과 축소본을  $p \times I$ 설계로 D 연구를 수행한 결과 완본의 경우 21문항을 평가했을 때 0.87로 적정 수준의 일반화가능도 계수인 0.80을 상회하였으며, 13문항에서도 적정 수준의 일반화가능도 계수에 도달하였다. 축소본의 경우 12문항을 평가했을 때 0.77로 적정 수준의 일반화가능도 계수에 미치지 못하였으며, 최소 15문항 이상에서 신뢰도가 적정 수준에 도달하였다. 둘째, 축소본을  $p \times (I:H)$ 설계로 D 연구를 수행한 결과 6영역에 대해 영역별로 2문항씩 구성될 경우 0.71로 적정 수준의 일반화가능도계수인 0.80 보다 낮게 측정되었으며, 최소 영역별 5문항 이상에서 신뢰도가 적정 수준에 도달하였다.

주요어: 집단논리적사고력검사, 일반화가능도이론, 신뢰도

\*Corresponding author: pioong2@hanmail.net

Tel: 82-43-261-2737

Fax: 82-43-271-0526

## 서 론

### 연구의 필요성과 목적

과학적 사고는 과학 지식의 형성과정에서 공통적으로 나타나는 사고의 특성이다. 이 과학적 사고에서 중추적인 역할을 하게 되는 사고를 논리적 사고라고 하며, 이 논리적 사고는 보존논리, 비례논리, 조합논리, 확률논리, 상관논리, 변인통제논리로 구분된다(Piaget and Inhelder, 1958; Popper, 1965; Lawson, 1995). Piaget and Inhelder(1958)는 문제 해결 상황에서 아동과 청소년이 문제를 해결하는 과정을 관찰하여 논리적 사고의 유형을 보존논리, 비례논리, 조합논리, 상관논리, 확률논리, 변인통제논리로 제시하였다. 그리고 Popper(1965)는 과학 지식의 발견과정에서 논리적 사고가 중추적인 역할을 함을 강조하고 있다. 즉, 논리적 사고력은 문제해결 과정에서 구체적으로 관찰할 수 있는 논리성을 내포한 사고를 의미한다.

논리적 사고력을 측정 하는 데는 여러 가지 도구를 사용하고 있는데 대체로 임상적면담법, 시범실험, 지필검사법 등으로 나눌 수 있다. 임상적면담법은 재생(reproducibility)이 어렵고 대량으로 투입하기가 어려우며, 시간이 많이 걸리고 일치된 점수를 얻기 위해서는 숙련된 면담자가 필요하다는 단점이 있다. 시범 실험 역시 면담법의 단점을 다소 보완하기는 하나 여전히 시간, 비 숙련된 면담자, 비 규격화된 방법 등의 문제점을 가지고 있다. GALT(Group Assessment of Logical Thinking; Roadrangka et al., 1983)는 이러한 단점을 보완하여 효율적이고 편리하게 연구자들이 큰 어려움 없이 집단적으로 인지수준을 검사 할 수 있는 지필검사법으로 논리적 사고력을 측정하는 도구로서 가장 널리 사용되고 있다.

GALT는 1983년 Vantipa Roadrangka, Russell H. Yeany, Michael J. Padilla가 공동으로 개발한 논리적 사고력 검사지이다. 총 문항은 21문항이고 각 문항은 답과 이유를 선택하는 선다형 형식으로 6개의 논리 유형을 측정하도록 구성되었다. 보존논리 4문항, 비례논리 6문항, 변인통제논리 4문항, 확률논리 2문항, 상관논리 2문항, 조합논리 3문항이고 대부분의 문항에 문제 상황을 설명하는 짧은 문장과 함께 그림을 제시하였고 6학년 이상의 학생들에게 1시간의 수업 시간 동안 집단으로 적용할 수 있다. 이 검사지의 타당도는 피아제의 임상적 면담법과의 상관관계로 나타

냈는데 조합논리가 0.88로 가장 높고 상관논리가 0.45로 가장 낮게 나타났으며 전체 타당도 계수는 0.71이다(Roadrangka et al., 1983).

GALT 개발자들은 시간 또는 다른 제약이 있을 경우 각 논리 유형별로 2문항씩을 뽑아 총 12문항으로 구성된 축소본 GALT의 사용을 제안하였다. 우리나라의 경우 원본 GALT는 최영준 외(1985)가 우리말로 번역하여 사용한 이후 많이 사용되어 오다가 1988년 이후로는 축소본 GALT(유갑렬, 1988)도 사용되어 현재는 과학 교육 연구에 병용되고 있으며, 점차 축소본 GALT의 사용이 증가하는 추세이다. GALT 축소본은 논리 유형별로 2문항씩 구성되어 있기에 각 논리유형들이 태도를 비롯한 과학에 관련된 또 다른 변인을 구성하는 요인과의 관계를 확인하는 연구에서 이용하기 용이하다. 다수의 선행연구들에서도 각 논리유형들과 학업성취도, 과학탐구능력, 과학과 관련된 태도와 같은 또 다른 요인들과의 관계를 확인하기 위해 GALT 축소본을 이용하였다(임정환과 정진우, 1991; 정연태와 김범기, 1991; 최재환, 1993; 소원주와 우종욱, 1994; 김태선 외, 2002).

신뢰도란 일반적으로 같은 피험자가 다른 상황에서 같은 시험을 다시 보았을 때 얻어지는 검사 점수들의 일관성(Anastasi, 1961)을 의미한다고 볼 수 있으나, 검사 점수의 신뢰도를 산출하기 위해서는 검사 점수를 구성하는 자료의 형태나 성격, 그리고 맥락에 따라 접근 방법이 달라져야 한다(양지승과 이규민, 2007). 검사의 신뢰도를 측정 시 검사 내에 다수의 요인이 존재하여 변인 간 분산이 복합적으로 존재하는 경우 신뢰도의 측정을 달리해야 하며(성태제, 2002), 일괄적으로 내적일관성 신뢰도(Cronbach's alpha)로 추정할 경우 검사 점수의 신뢰도가 과대 추정될 수 있다(이규민, 2003; 이용원, 2003; Thorndike, 1951; Anastasi, 1961; Sireci et al., 1991). 일반화가능도 이론(Generalizability theory; Cronbach et al., 1972)은 단일오차원(sources of error)만을 고려하는 고전검사 이론을 확대하여, 중다오차원(multiple sources of error)을 동시에 고려하는 측정모형에 분산분석(ANOVA) 체계를 적용한 이론으로 신뢰도에 영향을 미치는 다양한 요인들의 상호관계를 측정하는데 적합하다(성태제, 2002).

GALT는 그 이용 목적이 총 점수를 기준으로 인지수준을 구체적조작기, 과도기, 형식적조작기로 구분하여 사전과 사후의 측정 결과를 비교하는 연구, 그리

고 GALT를 구성하는 보존논리, 비례논리, 변인통제 논리, 확률논리, 상관논리, 조합논리를 세부적으로 조 명하고 나아가 이러한 하위 논리들이 다른 변인들과 의 작용하는 영향을 확인하는 연구로 구분된다고 할 수 있다. 이 때 전자의 경우 피험자의 인지수준 측정 결과에 영향을 미치는 국면이 문항요인 하나이지만, 후자의 경우 국면이 문항요인 외에도 문항을 포함하 는 영역 요인이 존재한다. 즉 GALT의 신뢰도를 측 정하는데 있어 이용 목적에 따라 산출방법을 다르게 적용할 필요가 있다. 즉, 본 연구의 목적은 GALT의 이용 목적에 따른 단일국면(single-facet)과 다국면 (multi-facet)의 상황에서 설계를 달리하여 오차원을 분석하고 신뢰도를 측정하는데 있다.

**연구 문제**

이상의 목적을 달성하기 위하여 본 연구에서 설정 한 연구문제는 다음과 같다.

연구문제 1. 원본과 축소본의 단일국면 설계(p×i) 시 오차원과 신뢰도는 어떠한가?

연구문제 2. 축소본의 다국면 내재설계(p×(i:h)) 시 오차원과 신뢰도는 어떠한가?

**연구방법 및 절차**

**연구대상**

본 연구는 지방 소재 초·중·고등학생을 대상으로 이루어졌다. 연구대상은 총 1016명으로 초등학교 5학 년이 131명(남학생: 74명, 여학생: 57명), 중학교 1학 년이 441명(남학생: 232명, 여학생: 209명), 그리고 고등학교 1학년이 444명(남학생: 186명, 여학생: 258 명)으로 구성되었으며, 이들을 대상으로 21문항의 GALT 원본을 40분 동안 실시하였다.

**검사도구**

본 연구에서는 Roadrangka et al.(1983)에 의하여 개발된 GALT 검사지를 최영준 외(1985)가 번역한 것을 보완하여 사용하였다. 총 21 문항으로 6개의 하 위 논리(보존, 비례, 변인통제, 확률논리, 상관논리,

**Table 1.** Elements of full version and short version

| Subtest                 | Item |                        | Version |       |
|-------------------------|------|------------------------|---------|-------|
|                         | No.  | Contents               | full    | short |
| Conservation            | 1    | Piece of Clay          | v       | v     |
|                         | 2    | Test Tube              | v       |       |
|                         | 3    | Road                   | v       |       |
|                         | 4    | Metal Weights          | v       | v     |
| Proportional Reasoning  | 5    | Plastic Jar #1         | v       |       |
|                         | 6    | Plastic Jar #2         | v       |       |
|                         | 7    | Glass Size #1          | v       |       |
|                         | 8    | Glass Size #2          | v       | v     |
|                         | 9    | Scale #1               | v       | v     |
|                         | 10   | Scale #2               | v       |       |
| Controlling Variables   | 11   | Pendulum Length        | v       |       |
|                         | 12   | Pendulum Weight        | v       | v     |
|                         | 13   | Ball #1                | v       | v     |
|                         | 14   | Ball #2                | v       |       |
| Probabilistic Reasoning | 15   | Square and Diamonds #1 | v       | v     |
|                         | 16   | Square and Diamonds #2 | v       | v     |
| Correlational Reasoning | 17   | The Mice               | v       | v     |
|                         | 18   | The Fish               | v       | v     |
| Combinatorial Reasoning | 19   | The Dance              | v       | v     |
|                         | 20   | The Shopping Center    | v       | v     |
|                         | 21   | Light Box              | v       |       |
|                         | N    |                        | 21      | 12    |

조합논리)로 구성되어 있으며, 이 중 축소본 검사는 별도로 실시하지 않고 이미 얻어진 원본의 자료에서 축소본에 해당하는 12문항을 추출하여 이용하였다 (Table 1).

### 분석방법 및 절차

일반화가능도 이론은 크게 일반화 연구(generalizability study, G 연구)와 결정 연구(decision study, D 연구)로 나누어진다. 분석의 절차는 먼저 측정 결과에 어떤 오차 요인이 얼마만큼 영향을 주는지 그 상대적 크기를 분석하기 위해 G 연구를 실시한다. 즉, 오차 요인에 따라 국면(facet)을 설정하고 자료 수집 형태가 교차(crossed)모형인지 내재(nested)모형인지 결정하여 분산분석(ANOVA) 설계를 적용한 후 분산분석 결과 얻어진 각 분산원의 제곱평균(MS)으로부터 분산성분(variance component)을 추정하여, 분산성분의 상대적 크기를 비교하여 각 오차원의 영향력을 분석한다. 그리고 논리적사고력 측정 결과를 얼마나 신뢰할 수 있는지 그리고 신뢰할 수 있는 평가가 되기 위해서는 어떤 조건을 갖추어야 하는지 알아내기 위하여 D 연구를 실시한다. 즉, G 연구에서 산출된 오차원의 분산성분을 토대로 일반화가능도 계수를 산출한 후 오차분산의 각 국면의 수를 늘림으로써 적정 수준(.80)의 일반화가능도 계수를 얻는 조건을 추정할 수 있다.

연구 분석의 자료 처리는 GENOVA(GENERalized analysis Of VAriance) 프로그램을 사용하였다. GENOVA는 Brennan and Lockwood(1980)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 and Brennan, 1983).

본 연구에서는 GALT의 이용 목적에 따라 단일국면과 다국면의 상황에서 설계를 달리하여 오차원을 분석하고 신뢰도를 측정하고자 한다. 먼저 단일국면 교차설계로 G 연구( $p \times i$  design)를 적용하고 이를 토대로 D 연구( $p \times I$  design)를 수행하고자 하며, 다국면 내재설계로 G 연구( $p \times (i:h)$ )를 적용하고 이를 토대로 D 연구( $p \times (I:H)$ )를 수행하고자 한다. 그러나 GENOVA 프로그램은 다국면의 상황에서 불균형 모델에 대한 일반화가능도 연구가 불가능하다(Crick and Brennan, 1983). 즉, GALT 축소본의 경우 각 논리유형을 구성

하는 문항이 동일하게 2문항으로 구성이 되어있으나 원본의 경우 각 논리유형의 문항이 불균형적이기 때문에 본 연구에서 다국면 내재설계를 통한 연구는 축소본에 대해서만 수행하였다.

### 분석의 이론적 근거

일반화가능도 계수(Generalizability coefficient,  $\rho^2$ )는 절대연구와 상대연구에 따라 다음과 같이 정의한다.

$$\rho^2(\delta) = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2} \quad \text{또는} \quad \rho^2(\Delta) = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2}$$

$\sigma_\delta^2$ : 상대오차분산: 관측점수와 전집점수 상대적 차이분산

$\sigma_\Delta^2$ : 절대오차분산: 관측점수와 전집점수 절대적 차이분산

잠재특성모형(latent trait model)에 기초하는 문항 반응이론(item response theory)에 병행하여 일반화가능도 이론은 무선표집모형(random sampling model)을 기본으로 발전하였다. 즉, 관찰된 단일 행위는 관찰행동 전집을 대표하는 행위로 보고 이러한 표집행위가 얼마나 일반화되어질 수 있는가 추정하는 과정이다. 일반화가능도 이론에서 전집(universe)은 모집단(population)의 의미와 약간 차이가 있다. 모집단은 표집단위가 측정대상인 반면, G이론에서 ‘전집’이란 ‘측정대상의 측정조건들에 대한 일반화과정’을 포함한다. 따라서 관찰 전집은 연구자의 목적과 실용적 관심을 토대로 오차요인을 이루는 하나 이상의 국면으로 구성된다. 여기서 측정의 ‘局面 (facet)’이란 ‘측정조건들의 집합’을 의미하며, 분산분석 설계에서 효과(effect)나 요인(factor)과 유사 개념이다. 또한 분산분석에서 처리(treatment)나 수준(level)은 조건(condition)으로 표현된다. G이론의 설계모형은 측정대상을 제외한, 즉 오차분산과 관련된 측정국면만을 연구 설계의 단위로 한다.

일반화가능도 이론을 적용한 연구는 언급한 바와 같이 두 단계로 진행된다. 첫째 허용 가능한 관찰전집을 정하여 G 연구 설계를 수행하여 분산성분 추정값을 산출하여 오차요인의 상대적 영향력을 파악하고, 둘째 D 연구 설계에 따라 전집점수 분산과 오차점수 분산을 정하여 일반화가능도 계수를 추정한다. 따라서, 일반화가능도 이론의 적용은 G 연구 설계에 따라 분산분석의 결과를 어떻게 활용할 수 있는가가

중요하다(김성숙과 김양분, 2001; Brennan and Lockwood, 1980; Shavelson et al., 1991; Brennan, 1992).

### 분석 결과 및 논의

본 연구에서는 고전검사이론을 이용한 GALT 원본과 축소본의 신뢰도를 측정하고 이용 목적에 따라 피험자의 인지수준 측정결과에 영향을 미치는 국면이 문항요인 하나인 경우 단일국면 설계(pxi), 검사의 영역 안에 다수의 국면이 포함되어 있는 경우 다국면 내제설계(p×(i:h))로 하여 오차원을 분석하였다. 그리고 각각의 국면 설계 시 적정 일반화가능도 계수를 지시하는 최적화된 조건을 추정하였다.

#### 고전검사이론을 이용한 GALT 원본과 축소본의 신뢰도 분석 결과

측정도구의 신뢰도의 수준에 관한 다양한 견해가 있으나 일반적으로 신뢰도 계수가 0.80 이상일 때 그 검사는 양호하다고 보고 있다(Dunbar et al., 1992; Shavelson et al., 1993). Cronbach's alpha 계수를 이용한 GALT의 신뢰도 측정결과 원본의 경우 0.87로 측정되었으며, 축소본의 경우 0.77로 측정되었다(Table 2).

이러한 신뢰도의 차이는 문항 수가 감소한 원인이 주요인이라고 추정되며, 더불어 축소본을 구성하는 문항들이 상대적으로 낮은 신뢰도를 갖고 있을 수도 있다고 볼 수 있다. 즉, 보다 명확한 규명을 위하여 피험자영향과 문항영향을 고려한 단일국면 설계를 통해 오차원을 분석할 필요가 있다.

**Table 2.** Reliability of GALT according to version

|              | Full version | Short version |
|--------------|--------------|---------------|
| N            | 1016         | 1016          |
| Items        | 21           | 12            |
| Cronbach's α | 0.87         | 0.77          |

**Table 3.** Result of G-study on full version (p×i design)

| Source of Variation | Degrees of Freedom | Sums of squares for score effects | Mean squares | Variance components for single observations(%) |
|---------------------|--------------------|-----------------------------------|--------------|--|
| Persons (p)         | 1015               | 1273.56                           | 1.25         | 0.052 (20.63)                                  |
| Items (i)           | 20                 | 730.15                            | 36.51        | 0.036 (14.29)                                  |
| Residual (pi, e)    | 20300              | 3325.09                           | 0.16         | 0.164 (65.08)                                  |
| Total               | 21335              | 5328.80                           |              | 0.252 (100.00)                                 |

#### GALT 원본과 축소본의 단일국면 설계 G 연구 결과(p×i design)

GALT의 충점을 기준으로 구체적조작기, 과도기, 형식적조작기로 구분하여 사전과 사후의 측정 결과를 비교하는 연구의 경우 오차원을 피험자요인(p)과 문항요인(i)으로 두고 단일국면 설계를 할 수 있으며, 이에 따른 원본과 축소본의 G 연구를 수행하였다.

Table 3은 GALT 원본 문항을 일반화가능도 이론을 적용하여 G 연구 결과로 산출된 각 분산 성분의 비율을 나타낸 것이다. 표에서 p는 전집분산이며, i와 pi, e는 오차분산에 해당된다. 전체 분산 중 피험자와 문항의 상호작용 성분(pi)과 오차(e)가 포함된 잔차(residual) 분산이 65.08%로 가장 큰 비율을 차지하는 것으로 나타났다. 학생간의 차이를 의미하는 피험자 분산(p)은 20.63%로 나타났으며, 문항간의 차이를 의미하는 문항분산(i)은 14.29%로 나타났다.

Table 4는 GALT 축소본 문항을 일반화가능도 이론을 적용하여 G 연구 결과로 산출된 각 분산 성분의 비율을 나타낸 것이다. 전체 분산 중 피험자와 문항의 상호작용 성분(pi)과 오차(e)가 포함된 잔차(residual) 분산이 가장 큰 비율을 차지하는 것으로 나타났다. 학생간의 차이를 의미하는 피험자분산(p)은 17.93%로 나타났으며, 문항간의 차이를 의미하는 문항분산(i)은 17.13%로 나타났다.

문항분산(i)이 차지하는 비율이 크다는 것은 문항간에 차이가 작아 문항의 특성이 학생들의 점수에 미치는 영향이 크다는 것으로 해석할 수 있으며, 피험자 분산(p)이 차지하는 비율이 크다는 것은 피험자의 특성이 점수에 미치는 영향이 크다는 것을 의미한다. 또한 잔차가 전체 분산 중 가장 크다는 것은 명확하게 설명될 수는 없지만 문항 국면 이외에 오차분산에 기여하는 또 다른 변동요인이 있을 수 있음을 암시한다. 일반적으로 피험자 분산이 클수록 일반화가능도 계수는 크게 산출되며, 문항 분산이 클수록 일반화가능도 계수는 작게 산출된다. 분석 결과 원본과 축소본의 경우 모두 잔차분산이 가장 크며,

**Table 4.** Result of G-study on short version (p×i design)

| Source of Variation | Degrees of Freedom | Sums of squares for score effects | Mean squares | Variance components for single observations (%) |
|---------------------|--------------------|-----------------------------------|--------------|---|
| Persons (p)         | 1015               | 719.63                            | 0.71         | 0.045(17.93)                                    |
| Items (i)           | 11                 | 479.91                            | 43.63        | 0.043(17.13)                                    |
| Residual (pi, e)    | 11165              | 1821.34                           | 0.16         | 0.163(64.94)                                    |
| Total               | 12191              | 3020.88                           |              | 0.251(100.00)                                   |

**Table 5.** Result of D-study on full version (p×I design)

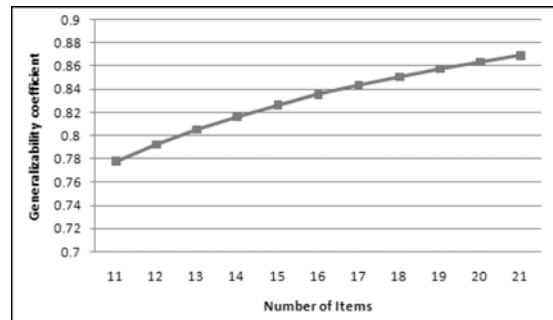
| Number of facet | N    | Item | Universe score | Expected observed score | Lower case delta | Upper case delta | Generalizability coefficient ( $\sigma^2$ ) | PHI   |
|-----------------|------|------|----------------|-------------------------|------------------|------------------|---|-------|
| Ni=1            | 1016 | 1    | 0.052          | 0.216                   | 0.164            | 0.200            | 0.24  | 0.207 |
| Ni G study      | 1016 | 21   | 0.052          | 0.060                   | 0.008            | 0.010            | 0.87  | 0.845 |
| Level 0.80      | 1016 | 13   | 0.052          | 0.065                   | 0.013            | 0.015            | 0.81  | 0.772 |

피험자분산이 문항분산에 비해 크게 나타났다. 즉, 문항 국면 이외에 오차분산에 기여하는 변동요인이 크게 작용하고 있다고 볼 수 있으며, 이러한 원인은 잠재적으로 내재되어 있는 하위 논리들을 구분 짓는 요인으로 인한 영향이라고 추정된다. 그리고 원본과 축소본 모두 피험자분산이 문항분산에 비해 크게 나타남은 문항 수에 의한 영향보다 피험자 특성인 피험자 간의 인지수준 차이가 더 큰 영향을 미치고 있다고 추정된다. 그러나 원본에서는 축소본 보다 피험자분산이 문항분산에 비해 더 큰 격차를 보이고 있으며, 축소본은 반대로 피험자분산과 문항분산이 근접해 있음으로 원본에 비해 축소본이 문항 수에 의한 영향을 더 크게 받고 있음을 알 수 있다.

**GALT 원본과 축소본의 단일국면 설계 D 연구 결과(p×I design)**

G 연구를 토대로 D 연구를 수행함으로써 국면의 수에 따른 일반화가능도 계수를 파악할 수 있으며, 최적화된 일반화가능도 계수를 나타내는 국면의 수를 예측할 수 있다. 일반적으로 G 연구에서 전집분산(피험자 분산)의 비율이 크고 오차분산(문항 분산+잔차 분산)의 비율이 작을수록 일반화가능도 계수는 크게 추정된다.

Table 5는 GALT 원본의 G 연구를 토대로 수행된 D 연구의 결과이다. D 연구 결과 GALT 원본은 21 문항을 평가했을 때 0.87로 적정 수준의 일반화가능도 계수인 0.80을 상회하였다. 그리고 13문항에서 최적화된 일반화가능도 계수인 0.81로 측정된 것으로 보아 21문항의 평가를 모두 할 필요 없이 13문항만



**Fig. 1.** Change of G-coefficient due to increase of item numbers on full version.

을 이용하여도 적정 수준의 신뢰도를 갖는 것으로 나타났다.

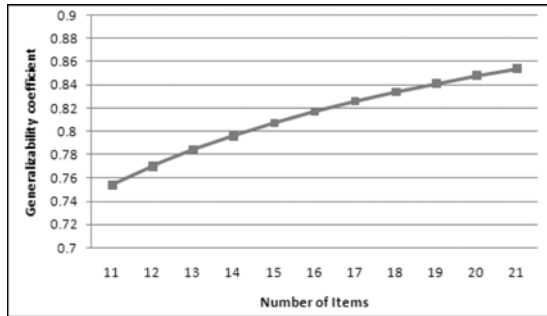
Fig. 1은 GALT 원본에 대한 문항 수 증가에 따른 일반화가능도계수의 변화 그래프로 13문항부터 일반화가능도 계수의 적정 수준인 0.80 이상임을 보여주고 있다.

Table 6은 GALT 축소본의 G 연구를 토대로 수행된 D 연구의 결과로 추정된 일반화가능도 계수를 나타낸 것이다. D 연구 결과, 일반화가능도 계수가 GALT 축소본의 경우 12문항만 할 경우 일반화가능도계수가 0.77으로 적정 수준의 일반화가능도계수인 0.80 보다 낮게 측정되었으며, 최소 15문항 이상에서 신뢰도가 적정 수준에 도달하였다.

Fig. 2는 GALT 축소본에 대한 문항수 증가에 따른 일반화가능도계수의 변화 그래프로 15문항부터 일반화가능도 계수의 적정 수준인 0.80 이상임을 보여주고 있다.

**Table 6.** Result of D-study on short version (p×I design)

| Number of facet | N    | Item | Universe score | Expected observed score | Lower case delta | Upper case delta | Generalizability coefficient ( $\sigma^2$ ) | PHI   |
|-----------------|------|------|----------------|-------------------------|------------------|------------------|---|-------|
| Ni=1            | 1016 | 1    | 0.045          | 0.209                   | 0.163            | 0.206            | 0.22  | 0.181 |
| Ni G study      | 1016 | 12   | 0.045          | 0.059                   | 0.017            | 0.004            | 0.77  | 0.726 |
| Level 0.80      | 1016 | 15   | 0.045          | 0.056                   | 0.011            | 0.014            | 0.81  | 0.768 |



**Fig. 2.** Change of G-coefficient due to increase of item numbers on short version.

GALT 원본과 축소본의 D 연구 결과 원본은 일반화가능도의 최적화 조건을 이루는 문항 수가 13문항이었으며, 축소본은 이보다 더 많은 15문항으로 측정

되었다. 이러한 결과는 원본과 축소본 간의 신뢰도의 차이가 문항 수 차이 외에도 신뢰도를 떨어뜨리는 또 다른 원인이 있음을 암시하며, 이는 원본에서 축소본의 문항을 구성하는 과정에서 상대적으로 신뢰도가 낮은 문항들의 비율이 높아졌기 때문이라고 추측된다.

GALT 원본과 축소본의 신뢰도 차이의 원인을 밝히고자 각 문항 제거 시 전체 평균, 분산, 그리고 신뢰도의 변화를 측정하였다. 확률논리와 상관논리는 기존의 2문항이 그대로 축소본으로 추출되었으므로 원본과 축소본의 신뢰도 차이에 영향을 주었다고 볼 수 없으며, 보존논리와 비례논리, 변인통제논리, 조합논리 역시 큰 영향을 주었다고 볼 수 없다. 그러나 상관논리의 경우 항목 전체 상관이 매우 낮은 문항들이며, 제거 시 문항의 신뢰도가 증가하는 문항들로

**Table 7.** Cronbach's alpha coefficient if item deleted from GALT (N=1016)

| Subtest                 | No. | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item-Total Correlation | Cronbach's $\alpha$ if Item Deleted |
|-------------------------|-----|----------------------------|--------------------------------|----------------------------------|-------------------------------------|
| Conservation            | 1*  | 9.373                      | 24.969                         | 0.305                            | 0.868                               |
|                         | 2   | 9.277                      | 25.243                         | 0.330                            | 0.868                               |
|                         | 3   | 9.507                      | 25.563                         | 0.118                            | 0.875                               |
|                         | 4*  | 9.764                      | 24.504                         | 0.329                            | 0.868                               |
| Proportional Reasoning  | 5   | 9.735                      | 23.298                         | 0.586                            | 0.859                               |
|                         | 6   | 9.743                      | 23.409                         | 0.563                            | 0.860                               |
|                         | 7   | 9.652                      | 23.023                         | 0.641                            | 0.857                               |
|                         | 8*  | 9.639                      | 23.029                         | 0.641                            | 0.857                               |
|                         | 9*  | 9.825                      | 23.908                         | 0.475                            | 0.863                               |
|                         | 10  | 9.689                      | 23.988                         | 0.431                            | 0.865                               |
| Controlling Variables   | 11  | 9.631                      | 23.031                         | 0.642                            | 0.857                               |
|                         | 12* | 9.606                      | 23.092                         | 0.632                            | 0.857                               |
|                         | 13* | 9.822                      | 24.141                         | 0.422                            | 0.865                               |
|                         | 14  | 9.821                      | 24.003                         | 0.453                            | 0.864                               |
| Probabilistic Reasoning | 15* | 9.645                      | 23.321                         | 0.576                            | 0.859                               |
|                         | 16* | 9.618                      | 23.227                         | 0.600                            | 0.859                               |
| Correlational Reasoning | 17* | 9.992                      | 26.146                         | 0.014                            | 0.876                               |
|                         | 18* | 10.114                     | 26.320                         | -0.010                           | 0.874                               |
| Combinatorial Reasoning | 19* | 9.471                      | 23.901                         | 0.500                            | 0.862                               |
|                         | 20* | 9.764                      | 23.416                         | 0.566                            | 0.860                               |
|                         | 21  | 9.758                      | 23.642                         | 0.514                            | 0.862                               |

\*Used items in short version

파악되었다. 즉, 문항 수는 12문항으로 줄었지만 상대적으로 17번과 18번 문항이 차지하는 비율이 증가하였기에 축소본의 신뢰도를 떨어뜨리는 요인으로 작용하였다고 추정된다(Table 7).

**GALT 축소본의 다국면 설계 G 연구 결과 (p×(i:h) design)**

인지수준을 구성하는 하위 요소인 보존논리, 비례논리, 변인통제논리, 확률논리, 상관논리, 그리고 조합논리의 영역으로 구분하는 연구의 경우 오차원을 피험자요인(p)과 문항요인(i), 그리고 영역요인(h)로 두고 다국면 설계를 할 수 있다. 이 설계는 검사 영역(h)에 따라 각각 다른 문항(i)으로 구성된 검사를 모든 피험자(p)에게 실시하여 검사 영역에 포함된 문항에 대한 피험자의 반응을 분석하는 설계이다. 본 설계에서 측정 대상은 피험자이며, 검사 영역과 검사 문항은 측정 국면이다. 또한 피험자, 검사 문항, 그리고 검사 영역은 무한대의 모집단과 전집으로부터 임의 표집 되었다고 가정하였다. 따라서 본 설계는 다국면 내재설계(multi-facet nested design)이며, 임의효과모형(random effect model)이다.

Table 8은 GALT 축소본 문항을 일반화가능도 이론을 적용하여 G 연구 결과로 산출된 각 분산 성분의 비율을 나타낸 것이다. 표에서 p는 전집분산으로 피험자간의 차이이며, 나머지는 오차분산으로 h는 검사 영역 간의 차이, i:h는 검사 영역 내 문항간의 차이, ph는 영역에 따라 피험자가 다르게 반응한 정도를 의미한다. 그리고 pi:h, e는 잔차에 의한 오차분산

이다. 피험자분산(p)은 16.67%로 나타났으며, 영역분산(h)은 6.75%로 나타났다. 검사 영역 내 문항분산(i:h)은 11.11%로 나타났고 피험자와 영역의 상호작용분산(ph)은 13.89%로 나타났다. 그리고 전체 분산 중 잔차분산이 51.59%로 가장 큰 비율을 차지하는 것으로 나타났다.

피험자분산(p)은 비교적 크게 나타나고 있는데 이는 검사 점수에 영향을 미치는 개인차의 정도가 크다는 것을 의미하며, 피험자들의 인지수준 차이가 크다는 것을 의미한다. 영역분산(h)은 비교적 작게 나타났다. 그러나 검사 영역 내 문항간 분산(i:h)과 영역에 따른 피험자분산(ph)이 비교적 높게 나타나고 있다. 이는 영역 간 검사 점수의 차이보다 영역내의 문항 간 점수의 차이와 각 영역에서 피험자의 검사 점수의 상대적인 위치의 차이가 더 크다는 것을 의미한다. 그리고 문항과 영역 국면 이외에도 여전히 오차분산에 기여하는 변동요인이 크게 작용하고 있으며, 보다 체계적으로 오차 분산을 감소시키기 위해서는 문항과 검사영역 이외의 오차원 분석이 필요하다고 보여진다.

**GALT 축소본의 다국면 설계 D 연구 결과 (p×(I:H) design)**

Table 9는 GALT 축소본의 다국면 내재설계 G 연구를 토대로 수행된 D 연구의 결과이다. D 연구 결과, 일반화가능도 계수가 영역별로 2문항씩 구성될 경우 0.71로 적정 수준의 일반화가능도계수인 0.80보다 낮게 측정되었으며, 최소 영역별 5문항 이상에서 신뢰도가 적정 수준에 도달하였다.

**Table 8.** Result of G-study on short version (p×(i:h) design)

| Source of Variation | Degrees of Freedom | Sums of squares for score effects | Mean squares | Variance components for single observations(%) |
|---------------------|--------------------|-----------------------------------|--------------|--|
| p                   | 1015               | 719.63                            | 0.71         | 0.042 (16.67)                                  |
| h                   | 5                  | 309.83                            | 61.97        | 0.017 ( 6.75)                                  |
| i:h                 | 6                  | 170.08                            | 28.35        | 0.028 (11.11)                                  |
| ph                  | 5075               | 1026.92                           | 0.20         | 0.035 (13.89)                                  |
| pi:h, e             | 6090               | 794.42                            | 0.13         | 0.130 (51.59)                                  |
| Total               | 12191              | 3020.88                           |              | 0.252 (100.00)                                 |

**Table 9.** Result of D-study on short version (p×(I:H) design)

| Number of facet (Nh=6) | N    | Item | Universe score | Expected observed score | Lower case delta | Upper case delta | Generalizability coefficient (σ <sup>2</sup> ) | PHI   |
|------------------------|------|------|----------------|-------------------------|------------------|------------------|--|-------|
| Ni=1                   | 1016 | 1    | 0.042          | 0.070                   | 0.028            | 0.035            | 0.60   | 0.546 |
| Ni G study             | 1016 | 2    | 0.042          | 0.059                   | 0.017            | 0.022            | 0.71   | 0.658 |
| Level 0.80             | 1016 | 5    | 0.045          | 0.053                   | 0.010            | 0.014            | 0.80   | 0.750 |



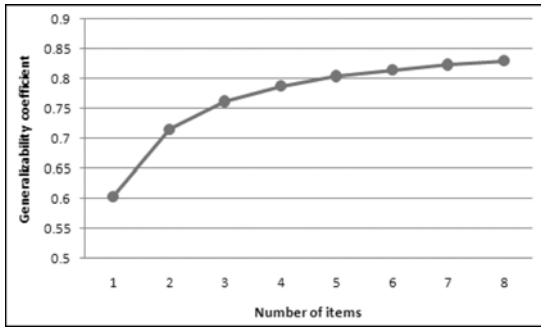


Fig. 3. Change of G-coefficient on short version (p x (I:H) design)

Fig. 3는 GALT 축소본의 다국면 내재설계로 D 연구를 통해 얻어진 일반화가능도 계수의 변화 그래프이다. 6요인으로 구성되는 GALT 축소본의 일반화가능도 계수는 현재의 영역별 2개 문항에서 일반화가능도계수가 0.71로 적정 수준에 도달하지 못하였으며, 최소 영역별 5개 문항 이상에서 일반화가능도계수가 적정 수준에 도달하고 있다.

이러한 결과는 학생들의 인지수준을 GALT의 총점을 이용하여 측정하고자 하는 연구가 아닌 학생들의 각 논리유형의 수준이 다른 변인들과의 작용하는 영향을 확인하는 연구에서 현재의 영역별 2문항으로 구성된 GALT 축소본은 적정 수준의 신뢰도에 도달하지 못하였으며, 보다 신뢰할 수 있는 측정을 위해서 각 논리 유형별로 최소 5개 이상의 문항을 개발하여 사용해야할 필요가 있다고 판단된다.

### 논 의

같은 피험자가 다른 상황에서 같은 시험을 다시 보았을 때 얻어지는 검사 점수들의 일관성을 의미한다(Anastasi, 1961). 고전검사이론을 이용한 GALT 원본과 축소본의 신뢰도 분석 결과 원본과 축소본간의 신뢰도는 차이가 있으며, 일관성이 부족하다고 할 수 있다. 이러한 신뢰도의 차이는 문항 수가 감소한 원인이 주요인이라고 판단된다.

GALT 원본과 축소본의 단일국면 설계로 G 연구를 수행한 결과 문항 국면 이외에 오차분산에 기여하는 변동요인이 크게 작용하고 있음을 확인하였으며, 이러한 원인은 잠재적으로 내재되어 있는 하위 논리들을 구분 짓는 요인으로 인한 영향이라고 판단된다. 그리고 원본과 축소본 모두 피험자분산이 문항

분산에 비해 크게 나타나고 있으며, 문항 수에 의한 영향보다 피험자 특성인 피험자 간의 인지수준 차이가 더 큰 영향을 미치고 있다고 판단된다. 그러나 원본에서는 축소본 보다 피험자분산이 문항분산에 비해 더 큰 격차를 보이고 있으며, 축소본은 반대로 피험자분산과 문항분산이 근접해 있음으로 원본에 비해 축소본이 문항 수에 의한 영향을 더 크게 받고 있다고 판단된다.

GALT 개발자들은 시간 또는 다른 제약이 있을 경우 각 논리 유형별로 2문항씩을 뽑아 총 12문항으로 구성된 축소본 GALT의 사용을 제안하였다(Roadrangka et al., 1983). 이 과정에서 축소본은 상대적으로 신뢰도가 낮게 나타난 상관논리의 문항 비율이 높아졌다. 이러한 이유 역시 원본에 비해 축소본의 신뢰도가 낮아지게 된 원인이라고 판단된다.

GALT 원본과 축소본의 단일국면 설계로 D 연구 결과 원본과 축소본 간의 신뢰도의 차이가 문항 수 차이 외에도 신뢰도를 떨어뜨리는 또 다른 원인이 있으며, 이는 원본에서 축소본의 문항을 구성하는 과정에서 상대적으로 신뢰도가 낮은 문항들의 비율이 높아졌기 때문이라고 추측된다. 상관논리의 경우 항목 전체 상관이 매우 낮은 문항들이며, 제거 시 문항의 신뢰도가 증가하는 문항들이기에 문항 수는 12문항으로 줄었지만 상대적으로 17번과 18번 문항이 차지하는 비율이 증가하여 축소본의 신뢰도를 떨어뜨리는 요인으로 작용하였다고 판단된다.

검사의 신뢰도를 측정 시 검사 내에 다수의 요인이 존재하여 변인 간 분산이 복합적으로 존재하는 경우 신뢰도의 측정을 달리해야 한다(성태제, 2002). 일반화가능도 이론(Generalizability theory; Cronbach et al., 1972)은 단일오차원(sources of error)만을 고려하는 고전검사이론을 확대하여, 중다오차원(multiple sources of error)을 동시에 고려하는 측정모형에 분산분석(ANOVA) 체계를 적용한 이론으로 신뢰도에 영향을 미치는 다양한 요인들의 상호관계를 측정하는데 적합하다(성태제, 2002).

논리 유형별로 각 2문항으로 구성된 축소본은 주로 하위 유형과 다른 변인들과의 관계를 파악하기 위한 연구에서 이용되고 있다(임청환과 정진우, 1991; 정연태와 김범기, 1991; 최재환, 1993; 소원주와 우종욱, 1994; 김태선 외, 2002). 이때 일괄적으로 내적일관성 신뢰도(Cronbach's alpha)로 추정할 경우 검사 점수의 신뢰도가 과대 추정될 수 있다(이규민, 2003; 이용원,

2003; Thorndike, 1951; Anastasi, 1961; Sireci et al., 1991). 일반화가능도 이론(Generalizability theory; Cronbach et al., 1972)은 단일오차원만을 고려하는 고전검사이론을 확대하여, 중다오차원을 동시에 고려하는 측정모형에 분산분석 체계를 적용한 이론으로 신뢰도에 영향을 미치는 다양한 요인들의 상호관계를 측정하는데 적합하다(성태제, 2002).

일반화가능도 이론을 이용하여 다국면 내재설계를 통한 G 연구 결과 영역 간 검사 점수의 차이보다 영역내의 문항 간 점수의 차이와 각 영역에서 피험자의 검사 점수의 상대적인 위치의 차이가 더 크며, 문항과 영역 국면 이외에도 여전히 오차분산에 기여하는 변동요인이 크게 작용하고 있다고 판단된다. 그리고 D 연구 결과 GALT를 구성하는 하위 논리의 영향을 고려하여 신뢰도를 추정할 경우 현재의 영역별 2문항으로 구성된 GALT 축소본은 적정 신뢰도를 갖지 못하며, 적어도 각 논리 유형별로 5개 이상의 문항을 개발하여 사용해야 할 필요가 있다고 판단된다.

## 결론

본 연구의 목적은 GALT 축소본의 이용 목적에 따라 일반화가능도 이론을 적용하여 문항과 피험자만을 고려한 단일국면과 문항과 피험자, 그리고 영역을 고려한 다국면의 오차원을 분석하는데 있다. 연구는 지방 소재 초·중·고등학교 총 1016명을 대상으로 이루어졌으며, 21문항의 GALT 완본을 수업시간 1시간 동안 실시한 후 이 중 축소본에 해당하는 12문항을 추출하여 일반화가능도를 이용한 신뢰도 분석에 이용하였다. 자료의 분석을 위해 고전검사이론을 이용한 GALT 완본과 축소본의 신뢰도를 분석하였으며, 일반화가능도 이론을 적용하여 단일국면 설계와 다국면 내재설계로 나누어 G 연구와 D 연구를 실시하였다.

고전검사이론을 이용한 GALT 완본과 축소본의 신뢰도를 Cronbach's alpha 계수를 이용하여 측정한 결과 완본의 경우 0.87로 측정되었으며, 축소본의 경우 0.77로 측정되었다. 완본과 축소본을 단일국면 설계(pxi design)로 G 연구를 수행한 결과 완본의 경우 전체 분산 중 잔차분산이 65.08%로 가장 큰 비율을 차지하는 것으로 나타났으며, 피험자분산(p)은 20.63%, 그리고 문항분산(i)은 14.29%로 나타났다. 그리고 축소본의 경우 잔차분산이 64.94%로 가장 큰 비율을 차지하는 것으로 나타났으며, 피험자분산(p)은 17.93%,

그리고 문항분산(p)은 17.13%로 나타났다. 이러한 결과는 완본과 축소본 모두 문항 국면 이외에 오차분산에 기여하는 변동요인이 크게 작용하고 있다고 해석할 수 있으며, 문항 수에 의한 영향보다 피험자 특성인 피험자 간의 인지수준 차이가 더 큰 영향을 미치고 있다고 보여진다. 그러나 완본에서는 축소본 보다 피험자분산이 문항분산에 비해 더 큰 격차를 보이고 있으며, 축소본은 반대로 피험자분산과 문항분산이 근접해 있음으로 완본에 비해 축소본이 문항 수에 의한 영향을 더 크게 받고 있음을 알 수 있다.

완본과 축소본을 단일국면 설계(p×I design)로 D 연구를 수행한 결과 완본의 경우 21문항을 평가했을 때 0.87로 적정 수준의 일반화가능도 계수인 0.80을 상회하였으며, 13문항에서 최적화된 일반화가능도 계수인 0.80으로 측정되었다. 축소본의 경우 12문항을 평가했을 때 0.77로 적정 수준의 일반화가능도 계수에 미치지 못하였으며, 최소 15문항 이상에서 신뢰도가 적정 수준에 도달하였다. 이러한 결과는 완본과 축소본 간의 신뢰도의 차이가 문항 수 차이 외에도 신뢰도를 떨어뜨리는 또 다른 원인이 있음을 암시하며, 이러한 원인은 각 문항 제거 시 전체 평균, 분산, 그리고 신뢰도의 변화를 측정한 결과 완본에서 축소본의 문항을 구성하는 과정에서 상대적으로 신뢰도가 낮은 문항들의 비율이 높아졌기 때문이라고 추정된다.

축소본을 다국면 내재설계(p×(i:h) design)로 G 연구를 수행한 결과 피험자분산은 16.67%로 나타났고 영역분산은 6.75%로 나타났으며, 검사 영역 내 문항분산은 11.11%로 나타났고 피험자와 영역의 상호작용분산은 13.89%로 나타났다. 그리고 잔차분산이 51.59%로 가장 큰 비율을 차지하는 것으로 나타났다. 이러한 결과는 피험자들의 인지수준 차이가 크다는 것을 의미하며, 영역 간 검사 점수의 차이보다 영역내의 문항 간 점수의 차이와 각 영역에서 피험자의 검사 점수의 상대적인 위치의 차이가 더 크다는 것을 의미한다. 축소본의 다국면 내재설계(p×(I:H) design)로 D 연구를 수행한 결과 영역별로 2문항씩 구성될 경우 일반화가능도계수는 0.71로 나타났으며, 영역별 최소 5개 이상의 문항이 되어야 적정 수준의 일반화가능도계수에 도달하였다. 단일국면 설계를 통해 얻어진 축소본의 신뢰도가 0.77이라는 것을 고려할 때 영역분산을 고려하여 측정한 GALT 축소본의 신뢰도는 보다 낮게 나타남을 확인하였다. 이러한 결과를 통해 GALT를 구성하는 하위 논리의 영향을 고려하여 신뢰

도를 추정할 경우 현재의 영역별 2문항으로 구성된 GALT 축소본은 적정 수준의 신뢰도에 도달하지 못하였으며, 학생들의 논리적 사고력을 구성하는 하위 논리의 수준을 확인하기 위해 GALT 축소본을 사용할 경우 각 논리 유형별로 최소 5개 이상의 문항을 개발하여 사용해야 할 필요가 있다고 판단된다.

연구결과를 토대로 GALT를 이용하여 학생들의 논리적 사고력을 측정하는 과정에서 보다 신뢰할 수 있는 측정을 위한 제언을 하고자 한다. 학생들의 논리수준을 측정하기 위해 축소본을 사용할 경우 적정 수준의 신뢰도를 얻지 못할 수 있으며, 보다 신뢰할 수 있는 측정을 위해 완본을 사용하여야 한다. 하위 논리유형인 보존논리, 비례논리, 변인통제논리, 확률논리, 상관논리, 조합논리를 확인하기 위해서는 논리 유형별로 5개 문항 이상으로 구성된 GALT를 새로이 개발할 필요가 있다. 더불어 본 연구는 한정된 지역과 인원을 대상으로 진행하였기에 전체를 대상으로 일반화하는데 어려움이 있으며, 보다 폭넓은 대상으로 심층적인 방법을 이용한 추가적인 연구가 진행되길 기대한다.

### 참고문헌

김성숙, 김양분, 2001, 일반화가능도 이론. 교육과학사, 서울, 121 p.  
 김태선, 배덕진, 김범기, 2002, 중학생의 그래프 능력과 논리적사고력 및 과학탐구능력의 관계. 한국과학교육학회지, 22, 725-739.  
 성태제, 2002, 타당도와 신뢰도. 학지사, 서울, 191 p.  
 소원주, 우종욱, 1994, 중학생의 논리적 사고력과 통합적 과학탐구 능력에 관한 한, 일 비교 연구. 한국과학교육학회지, 14, 312-320.  
 양지승, 이규민, 2007, 일반화가능도 이론을 적용한 단위검사 구성 검사점수의 신뢰도 추정. 교육평가연구, 20, 119-139.  
 유갑렬, 1988, 한국과 일본의 중학생의 과학적 사고력 및 탐구과정 능력에 관한 비교 연구. 서울대학교 석사학위논문, 102 p.  
 이규민, 2003, 단위검사 개념의 적용: 일반화가능도 이론을 중심으로. 교육평가연구, 16, 53-70.  
 이용원, 2003, Score reliability of a test composed of passage-based testlets. 영어교과교육, 2, 1-29.  
 임창환, 정진우, 1991, 고교생의 논리적사고력과 과학탐구기능 사이의 상관관계에 관한 연구. 한국과학교육학회지, 11, 23-30.  
 정연태, 김범기, 1991, 사고 훈련에 의한 우수 집단의 논리적 사고력 및 프로세스 스킬의 발달에 관한 연구. 과학교육논문집, 1, 241-247.  
 최영준, 최병순, 이원식, 1985, 중·고등학생들의 논리적 사고력 형성에 관한 연구(I). 한국과학교육학회지, 5, 1-9.

최재환, 1993, 뇌기능 분화가 논리적 사고력과 자연과 단원별 성취도에 미치는 영향. 과학·수학 교육연구, 17, 53-89.  
 Anastasi, A., 1961, Psychological testing. Macmillan, NY, USA, 657 p.  
 Brennan, R., 1992, Elements of generalizability theory. The American College Testing Program, IA, USA, 161 p.  
 Brennan, R.L. and Lockwood, R.E., 1980, A comparison of nedelsky and angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.  
 Bunce, D.M. and Hutchinson, K.D., 1993, The use of GALT(Group Assessment of Logical Thinking) as a predictor of academic success in college chemistry. Journal of Chemical Education, 70, 183-187.  
 Crick, J.E. and Brennan, R.L., 1983, Manual of GENOVA: A generalize analysis of variance system. American College Testing Program, IA, USA, 109 p.  
 Cronbach, L.J., Glesser, G.C., Nanda, H., and Rajaratnam, N., 1972, The dependability of behavioral measurements: Theory of generalizability for scores and profiles. John Wiley, NY, USA, 410 p.  
 Dunbar, S.B., Koretz, D.M., and Hoover, H.D., 1992, Quality control in the development and use of performance assessment. Applied Measurement in Education, 4, 289-303.  
 Lawson, A.E., 1995, Science teaching and the development of thinking. Wadsworth Publication, CA, USA, 593 p.  
 Piaget, J. and Inhelder, B., 1958, The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures. Basic Book, NY, USA, 356 p.  
 Popper, K.R., 1965, The logic of scientific discovery. Harper Torchbooks and The Science Library, NY, USA, 479 p.  
 Roadranga, V., Yeany, R.H., and Padilla, M.J., 1983, The construction and validation of group assesment of logical thinking (GALT). Paper Presented at the Annual Meeting of the National Association for Research in Science Teaching, 5 p.  
 Shavelson, R.J., Webb, N., and Burstein, L., 1991, Generalizability theory: A primer. Sage Productions, CA, USA, 152 p.  
 Shavelson, R.J., Baxter, G.P., and Gao, X., 1993, Sampling variability of performance assessments. Journal of Educational Measurement, 30, 215-235.  
 Sireci, S.G., Thissen, D., and Wainer, H., 1991, On the reliability of testlet-based tests. Journal of Educational Measurement, 28, 237-247.  
 Thondike, R.L., 1951, Reliability. In Lindquist, E.F. (ed.), Educational measurement. American Council on Education, Washington DC, USA, 560-620.

2009년 9월 1일 접수  
 2009년 10월 26일 수정원고 접수  
 2010년 1월 27일 채택