

## 정의적 성취특성 측정의 오차요인 및 성취수준별 척도의 동등성 분석<sup>1)</sup>

김 성 숙(한국교육과정평가원 선임연구위원)\*  
전 경 희(강남대학교 조교수)\*\*

### <요 약>

학업성취도 평가 결과 활용에 앞서 학생들의 다양한 배경변인, 특히 성취 수준에 관계없이 측정이 일관적이며 정확하게 이루어졌는지 점검할 필요가 있다. 이러한 맥락에서 본 연구의 목적은 일반화가능도 이론 모형을 적용하여 국가수준 학업성취도 평가의 정의적 영역에서의 성취수준별 측정의 동등성을 분석하는데 두었다. 이를 위해 2015년 중학교 3학년 학생 7,443명을 대상으로 실시한 학업성취도 평가 학생설문 자료 중 교과태도 관련 응답 결과에 대해 다변량 일반화가능도 분석을 실시하였다. G 연구에서는 정의적 특성의 측정 과정에서 발생하는 각 오차요인의 영향력을 분석하였으며, D 연구에서는 측정 조건 변화에 따른 신뢰도와 측정오차를 비교하였다. 분석 결과, 성취수준의 전 범위에서 걸쳐 교과태도 관련 정의적 성취 특성에 대한 측정이 대체로 양호한 것으로 나타났으나, 수학 교과에 대한 가치인식 영역에서 측정의 동등성을 담보하기에 다소 부정적인 결과를 보였다. 또한, 일부 측정 영역에서 성취수준과 교과에 따라 오차 요인들의 상대적 영향력이 다르게 나타났으며, 적정 수준의 신뢰도를 확보하기 위한 측정 조건의 변화 효과도 성취수준별로 상이하게 나타났다. 이러한 연구 결과를 토대로 다집단 일반화가능도 분석 결과가 피험자 집단별 측정의 동등성 여부를 점검하는 데 유용하게 사용될 수 있음을 확인하였다.

주제어 : 척도 동등성, 측정 불변성, 일반화가능도 이론, 정의적 성취특성

## I. 서론

최근 학교 교육에서 인성 및 핵심 역량 관련 교육을 비롯하여 정의적 영역에 대한 관심이

1) 이 논문은 ‘제9회 국가단위 평가 자료 분석 및 맞춤형 학력진단검사 활용 연구 세미나’ 발표 자료(전 경희, 김성숙, 2017)의 내용을 수정·보완한 것임.

\* 제1저자, sungs@kice.re.kr

\*\* 교신저자, kjeon@kangnam.ac.kr

높아짐에 따라 국가수준 학업성취도 평가(이하 학업성취도 평가)에서도 정의적 영역에 대한 측정의 중요성이 강조되고 있다. 이에 따라 학업성취도 평가에서는 국가수준 및 학교 교육의 책무성 차원에서 학생 설문 응답 결과에 나타난 우리나라 학생들의 정의적 영역에 대한 성취 특성을 지속적으로 점검하고 있다. 정의적 성취 특성은 인간의 여러 가지 행동 특성 중 정서, 감정, 느낌이 반영된 행동 특성을 의미하는 것으로 인지적 성취 특성과 대비되는 개념으로 간주된다. 인지적 성취 특성에 대한 측정도구와는 달리 정의적 성취 특성은 자기보고식 설문도구로 수집되는 경우가 일반적이다. 자기보고식 측정도구의 특성 상 피험자의 배경변인 또는 성향에 의해 응답 결과가 왜곡되거나 달라지는 문제가 발생할 수 있으며, 이와 같은 문제가 체계적인 방식으로 개입될 때 측정오차가 커지면서 측정의 정확성 또는 신뢰도에 부정적인 영향을 초래하게 된다. 실제 다양한 문화권의 응답자료 특성을 비교한 연구(Buckley, 2009; Mottus, Allik & Realo, 2012)에 따르면, PISA를 포함한 국제비교연구 자료에서 국가, 인종, 성별 등 피험자의 배경변인에 따라 응답양식이 상이하게 나타나고 있음을 경험적으로 확인함으로써 평가결과에 응답편파 효과가 개입되었음을 시사하였다.

한편, 동일한 국가 또는 문화권 내에서도 이와 같은 응답편파 또는 측정의 오차가 발생할 수 있으며, 학업성취도 결과 활용에 앞서 학생의 배경변인 및 특성에 관계없이 측정이 일관적이며 정확하게 이루어졌는지에 대한 점검이 요구된다. 즉, 학업성취도 평가의 양호도를 확보하기 위해 측정학적 이론을 기반으로 하여 피험자 집단별 척도의 불변성(the invariance of scales across groups) 또는 집단 불변성(group invariance) 개념이 유지되는지 확인할 필요가 있다. 특히, 최근 학교 교육에서 학생의 전인적 성장을 강조하며 인지적 영역과 정의적 영역에서의 조화로운 발달을 도모하는 맥락에서 학생들의 인지적 성취수준에 영향을 받지 않고 정의적 영역에 대한 측정이 정확하고 신뢰롭게 이루어졌는지의 여부를 확인하는 과정이 중요하게 인식되고 있다.

이에 본 연구에서는 일반화가능도 이론 모형을 적용하여 신뢰도 관점에서 학업성취도 평가의 정의적 영역에서의 성취수준별 측정의 동등성(measurement equivalence)을 검증하고자 한다. 이를 위한 구체적인 연구문제는 다음과 같다. 첫째, 국어, 수학, 영어 교과태도 관련 정의적 성취 특성의 측정 결과에 영향을 미치는 측정 오차 요인들은 무엇이며 그 영향력은 어느 정도인가? 둘째, 각 교과에서 성취수준이 다른 집단 간 신뢰도, 오차 분산, 및 조건부 측정 오차가 어떻게 다르게 나타나는가? 셋째, 적정 수준의 신뢰도에 도달하기 위해 정의적 특성의 측정구인별 문항 수는 어느 정도가 적절한가? 이러한 연구문제에 대한 실증적 자료 분석 결과를 기반으로 하여 우리나라 학생들의 정의적 성취 특성을 보다 정확하고 신뢰롭게 파악하는 데 기여하고자 한다.

## II. 이론적 배경

### 1. 정의적 성취 특성의 측정

정의적 성취 특성은 일반적으로 인지적 성취 특성과 대비되는 용어로 정의된다. 그러나, Popham(2003)과 OECD(2013, 2014)를 비롯한 다수의 선행연구에서 논의된 바와 같이 인지적 성취와 정의적 성취 변인 간 밀접한 관련성이 입증되고 있다. 학생의 미래 행동을 예측하는 맥락에서도 정의적 성취 특성이 매우 중요한 변인으로 작용하고 있음을 확인할 수 있다. 예를 들어, 자아개념, 수업에 대한 관심, 교사에 대한 만족도 등 정의적 영역에서의 성취가 우수한 경우 학업성취도 역시 높게 관찰되거나 전반적인 학교생활에 긍정적인 영향을 미치게 된다. 따라서 학생들의 정의적 특성을 함양하기 위한 노력은 학교 교육의 질을 높이는 문제와 밀접한 관련을 맺는다고 볼 수 있다.

최근 교육과정 및 학생평가 패러다임의 주요한 변화 중 하나는 인지적 영역에 비해 상대적으로 간과되어 온 정의적 영역에서의 성취 특성에 대한 측정의 중요성이 새롭게 부각되고 있다는 점이다. 실제 우리나라의 학업성취도 평가에서 뿐 아니라 PISA, TIMSS, ICILS와 같은 국제학업성취도 연구에서도 매 평가 주기마다 인지적 영역에 대한 성취도를 평가함과 동시에 별도의 측정도구를 이용하여 학생의 인지적 성취에 영향을 미치는 다양한 교육 맥락 변인들을 조사하고 있으나, 최근 들어 그 자료수집의 범위와 방법이 이전 주기에 비해 훨씬 심층적이며 정교하게 이루어지고 있는 추세이다. 그러나, 학업성취도 평가 연구를 비롯한 국내 연구결과에서 뿐 아니라 PISA와 TIMSS 등 국제학업성취도 평가를 통해 도출된 최근 연구 결과에 따르면, 우리나라 학생들은 인지적 영역에서 국제적으로 높은 성취 수준을 보이는 반면, 흥미, 자아효능감, 학습태도, 가치인식 등 정의적 영역에서의 성취는 매우 낮은 것으로 보고되고 있다(김경희, 김수진, 2010; 김수진 외, 2012; 최승현, 박상욱, 홍혜정, 2014). 따라서 우리나라 학생들의 인지적 영역에서의 성취도 뿐 아니라 정의적 영역의 성취 특성을 보다 분석적으로 살펴볼 필요가 있으며, 두 영역 간 균형 있는 교육을 실현하기 위한 구체적이고 실질적인 방안을 모색하기 위한 노력이 필요하다.

이러한 맥락에서 학업성취도 평가에서는 우리나라 초, 중, 고 학생들의 학습 관련 정의적 성취 특성을 파악하기 위해 수학, 국어, 영어 교과와 태도 관련 문항들에 대한 학생들의 응답 결과를 주기적으로 분석하고 있다. 이러한 교과태도 관련 변인에 대한 선행연구를 살펴보면, 특정 교과에 대한 흥미, 가치, 자신감 등에 대한 측정을 다루고 있다. 그 대표적인 예로 이진향

(1994)은 자신감, 유용성 지각, 흥미의 세 영역으로 분류한 후 수학 교과태도를 측정하였으며, 김재철(2002)은 필요성 인식, 자신감, 흥미의 세 영역으로 구분하여 교과에 대한 태도를 측정하였다. 학업성취도 평가 설문에서는 이러한 교과태도 측정을 위해, 2016년 이전 주기까지는 전수 설문에서 학교 급에 따라 학업성취도 평가 시행 교과별 학생들의 학습 태도를 묻는 문항이 교과 흥미, 교과 가치로 구성되었으며, 표집 설문을 통해 보다 다양한 범위에 걸쳐 교과태도에 대한 측정이 이루어져 왔다. 그러나, 2016년에는 기존 주기 표집 설문에 문항을 부분적으로 추가하여 자신감, 흥미, 가치, 학습의욕의 4개 하위요인으로 구성되는 수학 교과태도 척도를 개발하였다. 2016년 새롭게 개발된 척도에서 자신감은 ‘과제를 성공적으로 수행하거나 성취할 수 있는 자신의 능력에 대한 긍정적인 판단’으로 정의하고 있으며, 흥미는 ‘과제와 학습 활동에 대한 관심과 선호도 및 학습 활동을 수행하면서 경험하는 즐거움’을 의미한다(박인용 외, 2016). 또한 가치는 ‘학문적, 직업적, 사회적 맥락이나 학생의 삶의 맥락에서 각 교과와 기능과 유용성, 중요성에 대한 판단’을 나타내며, 학습의욕은 ‘학습하려는 인내와 노력 및 학습 상황에서 어렵고 낯선 문제나 과제에 도전하려는 자세’로 정의하고 있다(박인용 외, 2016).

## 2. 일반화가능도 이론의 적용

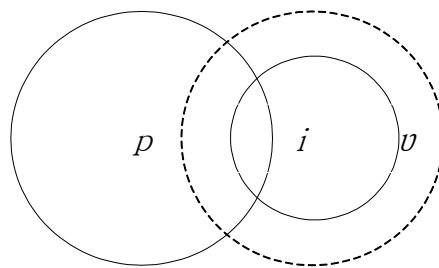
일반화가능도 분석이란 연구자가 설정한 특정 상황에서 발생하는 측정 오차의 원인을 세분화하여 피험자 점수 분산에서 각각의 오차가 차지하는 상대적인 비중을 파악함으로써 정확하고 효율적인 연구를 설계하는 데 유용한 정보를 제공하는 분석 방법이다(Brennan, 2001a, 이현숙, 2012). 이 연구에서는 학업성취도 평가 정의적 영역 중 교과태도와 관련한 학생들의 설문 응답결과를 분석하기 위해 일반화가능도 이론에 의한 방법을 적용하였다. 일반화가능도 이론에 의한 분석은 일반적으로 G 연구와 D 연구 단계로 진행되며, G 연구에서는 연구자가 관심을 갖는 모든 측정 조건들을 포함하는 허용 가능한 관찰 전집을 설정하고, D 연구에서는 어떠한 요인들로 측정 절차를 반복할 지를 구성하는 일반화 전집을 결정하게 된다(Brennan, 2001a, Lakin & Lai, 2012). 일반화가능도 분석의 용어인 전집점수(universe score), 국면(facet), 조건은 고전검사이론에서 각각 진점수(true score), 분산분석에서의 요인, 수준으로 간주할 수 있다. 또한 고전검사이론에서의 신뢰도계수가 상대평가에만 적용할 수 있는 반면 일반화가능도 분석에서는 상대평가와 절대평가 상황에 해당하는 신뢰도계수를 각기 다르게 산출한다. 즉, 상대평가 상황에서는 상대오차 분산을 활용하여 일반화가능도 계수가 산출되며, 절대평가 상황에서는 절대오차 분산을 활용한 의존도계수를 산출한다.

일반화가능도 모형을 적용한 연구 대부분이 단변량(univariate) 일반화가능도 모형에 의한 분석을 수행하였던 반면, Li와 Brennan(2007), Powers와 Brennan(2009), Yin(2005), 김성숙, 송미영, 박인용(2012), 이현숙(2012) 등은 다변량(multivariate) 일반화가능도 모형을 적용하여 다

양한 측정학적 맥락에서 검사 자료를 분석하였다. 단변량 일반화가능도 분석의 확장된 형태인 다변량 일반화가능도 분석은 모든 피험자가 고정국면의 각 조건에 해당하는 두 개 이상의 전집점수를 가정하면서, 각 점수는 상호 관련 있는 측정 영역 중 하나에만 대응되는 검사 자료에 적용할 수 있는 분석방법이다. 단변량 일반화가능도 분석과는 달리 다변량 일반화가능도 분석은 공분산 행렬을 제공함으로써 측정 영역별 전집점수에 가중치를 준 합성점수에 대한 오차 분산 및 신뢰도 산출을 가능하게 한다(Brennan, 2001a).

### 3. 다변량 일반화가능도 분석 설계

본 연구에서 적용한 다변량 일반화가능도 모형은  $p^{\bullet} \times i^{\circ}$ 로서, 일반적으로  $p^{\bullet} \times i^{\circ}$  설계에서 피험자(p)에 대한 닫힌 원은 해당 국면이 고정국면인 측정 요소(v)와 교차됨을 의미하며, 문항(i)에 대해 열린 원은 해당 국면이 고정국면인 측정 요소(v)와 내재되어 있음을 의미한다. 이 때, 각 측정 요소 수준별로 구분하여 분석이 이루어지므로 다변량 일반화가능도 분석의 기호 체계에서 측정 요소 국면을 의미하는 v는 별도로 표시되지 않는 것이 일반적이다. 즉, 모든 문항(i)이 모든 피험자(p)에게 실시되었으며, 세 개의 측정 요소(v)로 구성된 교과태도 관련 정의적 성취 특성에 대한 측정 상황을 가정하였다. 이 때, 검사 도구는 세 가지 측정구인만으로 구성됨을 가정하므로, 측정구인(v) 국면은 자신감, 흥미, 가치인식의 세 가지 수준으로만 한정된 고정국면으로 간주된다. 반면, 검사에 포함된 문항(i)과 피험자(p)는 무한 전집을 대표하도록 표집 되었으므로 임의국면으로 정의된다.  $p^{\bullet} \times i^{\circ}$ 에 대응하는 단변량 일반화가능도 분석 모형은  $p \times (i:v)$ 로서 벤다이어그램으로 표현하면 [그림 II-1]과 같다. 이러한  $p^{\bullet} \times i^{\circ}$  모형은 검사도구 설계 과정에서 각 문항이 이원분류표에 근거하여 특정 행동영역이나 내용영역의 요소를 측정하도록 출제되는 자료 구조와 대체로 일치하므로, Brennan(2001a)은 이 모형을 검사상세화표 모형(the table of specification model)으로 지칭하였다.



[그림 II-1]  $p^{\bullet} \times i^{\circ}$  설계의 벤다이어그램

$p^{\bullet} \times i^{\circ}$ 모형을 적용한 G 연구 설계에서 v 국면의 세 측정구인( $v_1$ ,  $v_2$ ,  $v_3$ )에 대한 피험자 점

수는 다음과 같이 표현된다.

$$X_{piv_1} = \mu_{v_1} + \nu_p + \nu_i + \nu_{pi} \quad (1)$$

$$X_{piv_2} = \mu_{v_2} + \nu_p + \nu_i + \nu_{pi} \quad (2)$$

$$X_{piv_3} = \mu_{v_3} + \nu_p + \nu_i + \nu_{pi} \quad (3)$$

식 (1)은 첫 번째 측정구인인 자신감( $v_1$ )에 대한 피험자 점수  $X_{\pi v_1}$ 가 피험자 효과( $\nu_p$ ), 문항 효과( $\nu_i$ ), 피험자와 문항의 상호작용 또는 잔차 효과( $\nu_{pi}$ ), 그리고 해당 측정 요소에 대한 전체 평균으로 구성되어 있음을 의미한다. 이와 마찬가지로 식 (2)와 (3)은 각각 흥미( $v_2$ )와 가치인식( $v_3$ ) 측정구인에 대한 피험자 점수의 선형 방정식을 나타낸다. G 연구 설계에서 임의국면에 대한  $v$  국면의 각 수준별 분산 및 공분산 행렬은 식 (4)-(6)과 같다. 아래 행렬식에서 대각선 위에 놓인 값은  $v$  국면의 분산 성분을 나타내고 대각선 밖의 원소는  $v$ 국면의 각 수준 간 공분산 성분을 의미한다.

$$\Sigma_p = \begin{bmatrix} \sigma_{v_1}^2(p) & \sigma_{v_1 v_2}(p) & \sigma_{v_1 v_3}(p) \\ \sigma_{v_2 v_1}(p) & \sigma_{v_2}^2(p) & \sigma_{v_2 v_3}(p) \\ \sigma_{v_3 v_1}(p) & \sigma_{v_3 v_2}(p) & \sigma_{v_3}^2(p) \end{bmatrix} \quad (4)$$

$$\Sigma_i = \begin{bmatrix} \sigma_{v_1}^2(i) & & \\ & \sigma_{v_2}^2(i) & \\ & & \sigma_{v_3}^2(i) \end{bmatrix} \quad (5)$$

$$\Sigma_{pi} = \begin{bmatrix} \sigma_{v_1}^2(pi) & & \\ & \sigma_{v_2}^2(pi) & \\ & & \sigma_{v_3}^2(pi) \end{bmatrix} \quad (6)$$

G 연구에서  $v$  국면(측정구인)의 각 수준별로 분산 및 공분산 성분이 추정되면 연구자에 의해 설정된  $v$  국면의 수준별 표본크기에 따라 D 연구가 수행된다. 이 때 G 연구에서 산출된  $\widehat{\Sigma}_p$  행렬은 D 연구에서도 동일하게 적용되며,  $\widehat{\Sigma}_i$ 와  $\widehat{\Sigma}_{pi}$  행렬을 D 연구의 표본크기로 각각 나누어  $\widehat{\Sigma}_I$ 와  $\widehat{\Sigma}_{pI}$  행렬이 추정된다. 이 값들을 이용하여 상대오차 분산과 절대오차 분산을 식

(7)-(8)과 같이 산출할 수 있다. 아래 식을 이용하여 상대오차 분산과 절대오차 분산이 추정된 후, 전체 분산 중 전집점수( $\widehat{\Sigma}_p$ )가 차지하는 비율로 고전검사이론의 신뢰도에 해당하는 일반화 가능성도 계수( $E\rho^2$ )와 의존도계수( $\widehat{\Phi}$ )를 산출할 수 있다.

$$\widehat{\Sigma}_\delta = \widehat{\Sigma}_{pI} \quad (7)$$

$$\widehat{\Sigma}_\Delta = \widehat{\Sigma}_I + \widehat{\Sigma}_{pI} \quad (8)$$

### III. 연구 방법

#### 1. 연구 자료

이 연구에서는 학업성취도 평가에 대한 2015년 중학교 3학년 표집자료로부터 국어, 수학, 영어 각 교과에 대한 자신감 4문항, 흥미 4문항, 가치인식 4문항을 포함하여 총 12개 교과태도 관련 변인에 대한 4점 리커트 척도 응답 결과를 분석하였으며, 각 교과별 교과태도 관련 정의적 특성의 측정구인 및 문항은 <표 III-1>과 같다.

<표 III-1> 교과태도 관련 정의적 특성 측정구인 및 설문 문항

교과태도	측정구인	설문 문항
국어 태도	자신감	<ul style="list-style-type: none"> <li>· 나는 다른 사람보다 국어를 잘하는 편이다</li> <li>· 나는 글을 읽고 의미를 정확하게 이해할 수 있다</li> <li>· 나는 다른 사람의 말을 듣고 요점을 잘 파악할 수 있다</li> <li>· 나는 생각을 글이나 말로 조리 있게 표현할 수 있다</li> </ul>
	흥미	<ul style="list-style-type: none"> <li>· 나는 국어를 공부하는 것이 즐겁다</li> <li>· 나는 국어 공부에 흥미가 있다</li> <li>· 나는 스스로 읽을거리를 찾아 읽는 편이다</li> <li>· 나는 글쓰기를 좋아한다</li> </ul>
	가치인식	<ul style="list-style-type: none"> <li>· 국어 공부는 내가 나중에 하고 싶은 일을 하는 데 도움이 될 것이다</li> <li>· 나는 다른 교과를 배우는 데 국어가 도움이 된다고 생각한다</li> <li>· 국어 공부는 우리말과 글의 소중함을 깨닫게 해 준다</li> <li>· 국어 공부는 내 생각을 조리 있게 표현하는 데 도움을 준다</li> </ul>
수학 태도	자신감	<ul style="list-style-type: none"> <li>· 나는 다른 사람보다 수학을 잘하는 편이다</li> <li>· 나는 친구에게 수학 공식을 설명해 줄 수 있다</li> <li>· 나는 지금보다 어려운 수학 문제도 풀 수 있다</li> <li>· 노력해도 나에게 수학은 여전히 어렵다</li> </ul>
	흥미	<ul style="list-style-type: none"> <li>· 나는 수학을 공부하는 것이 즐겁다</li> <li>· 나는 수학 공부에 흥미가 있다</li> <li>· 나는 금방 답이 나오지 않는 수학 문제를 푸는 것을 좋아한다</li> <li>· 나는 수(number)를 다루는 것을 좋아한다</li> </ul>

교과태도	측정구인	설문 문항
영어 태도	가치인식	<ul style="list-style-type: none"> <li>· 수학 공부는 내가 나중에 하고 싶은 일을 하는 데 도움이 될 것이다</li> <li>· 나는 다른 교과를 배우는 데 수학이 도움이 된다고 생각한다</li> <li>· 수학을 배우면 논리적으로 사고하는 데 도움이 된다</li> <li>· 일상생활에서 수학 지식이 꼭 필요한 것은 아니다</li> </ul>
	자신감	<ul style="list-style-type: none"> <li>· 나는 다른 사람보다 영어를 잘하는 편이다</li> <li>· 나는 내 생각을 영어로 표현할 수 있다</li> <li>· 나는 영어로 된 글을 읽고 이해할 수 있다</li> <li>· 나는 영어로 말하는 것을 듣고 요점을 파악할 수 있다</li> </ul>
	흥미	<ul style="list-style-type: none"> <li>· 나는 영어를 공부하는 것이 즐겁다</li> <li>· 나는 영어 공부에 흥미가 있다</li> <li>· 나는 영어로 된 읽을거리를 스스로 찾아 읽는 편이다</li> <li>· 나는 영어로 글쓰기를 좋아한다</li> </ul>
	가치인식	<ul style="list-style-type: none"> <li>· 영어 공부는 내가 나중에 하고 싶은 일을 하는 데 도움이 될 것이다</li> <li>· 나는 다른 교과를 배우는 데 영어가 도움이 된다고 생각한다</li> <li>· 영어 공부를 하면 다른 나라 문화를 이해하는 데 도움을 준다</li> <li>· 현대 사회에서 필요한 정보를 얻으려면 영어를 알아야 한다</li> </ul>

\* 2015년 중학교 3학년 표본에서 학습태도 관련 문항은 각 교과별로 총 14문항으로 구성되어 있으나, 분석 자료의 타당도 검증을 위한 요인분석 실시 결과, 14개 문항 중 2개 문항이 자신감, 흥미, 가치인식의 어느 요인에도 속하지 않은 별도의 문항 군으로 분류되거나 요인 부하량이 .4 이하로 매우 낮게 나타나 일반화가능도 분석에서 제외됨.

## 2. 다변량 일반화가능도 분석

본 연구에서는 교과태도 관련 정의적 성취특성 측정의 오차요인 및 성취수준별 척도의 동등성 분석을 위해 다변량 일반화가능도 모형을 적용한 G 연구와 D 연구를 수행하였다. 구체적으로  $p^* \times i^o$  설계를 적용하여 모든 문항( $i$ )이 모든 피험자( $p$ )에게 실시되었으며, 세 개의 측정구인( $v$ )으로 구성된 교과태도 관련 정의적 성취 특성에 대한 측정 상황을 반영하였다. 이 때, 검사 도구는 세 가지 측정구인만으로 구성되었음을 가정하므로, 측정구인( $v$ ) 국면은 자신감, 흥미, 가치인식의 세 가지 수준으로만 한정된 고정국면으로 간주하였다. 반면, 검사에 포함된 문항( $i$ )와 피험자( $p$ )는 무한 집합을 대표하여 표집된 것으로 가정하여 임의국면으로 설정하였다. 한편, 일반화가능도 분석에서 측정의 대상이 어떤 특성에 의해 총화되어 있다면 그 집단을 구분하여 분석하는 것이 바람직하다는 Brennan(2001a)의 제안에 따라 각 교과의 학업성취도 결과를 기준으로 성취수준별로 학생 집단을 ‘우수’, ‘보통’, ‘기초’, ‘기초미달’을 포함한 네 집단으로 구분하여 분석하였다. G 연구에서는 정의적 특성의 측정과정에서 발생하는 다중오차요인과 각 오차요인의 영향력을 분석하였으며, 성취수준 집단별 G 연구 결과를 비교하였다. D 연구에서는 측정구인별로 문항 수를 4개 문항에서 8개 문항으로 증가함에 따른 측정조건의 변화가 신뢰도와 측정오차에 미치는 영향력을 비교하였다. 이상에서 설명한 다변량 일반화가능도 분석을 위해 mGENOVA(Brennan, 2001b) 프로그램을 이용하였다.



## IV. 연구 결과

### 1. 성취수준별 교과태도 기술통계

일반화가능도 분석에 앞서 국어, 수학, 영어 세 교과에서 학습태도 관련 정의적 성취 특성을 성취수준별 차이를 분석한 결과는 <표 IV-1>과 같다. 분산분석 결과, 세 교과태도에 대한 모든 측정 영역에서 성취수준별 집단 평균에 유의한 차이를 보이는 것으로 확인되었다. 또한, 세 교과에서 공통적으로 성취수준이 높은 집단일수록 자신감, 흥미, 가치인식에 대한 정의적 성취 특성의 평균 점수가 대체로 높게 나타났다.

<표 VI-1> 성취수준별 교과태도 관련 정의적 성취특성의 기술통계 및 집단 간 차이 분석 결과

교과	측정구인	우수		보통		기초		기초미달		전체		F
		평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차	
국어	자신감	2.95	.55	2.69	.54	2.48	.62	2.34	.73	2.73	.59	213.23*
	흥미	2.74	.67	2.55	.63	2.43	.68	2.34	.72	2.58	.66	75.23*
	가치인식	3.12	.56	2.90	.59	2.71	.68	2.48	.80	2.93	.62	161.95*
수학	자신감	3.06	.58	2.56	.59	2.09	.61	2.03	.63	2.49	.69	816.28*
	흥미	2.93	.73	2.53	.74	2.10	.76	2.01	.80	2.46	.80	385.45*
	가치인식	2.96	.58	2.65	.59	2.40	.62	2.38	.54	2.62	.63	265.27*
영어	자신감	3.06	.58	2.57	.65	2.15	.74	1.96	.81	2.57	.75	657.04*
	흥미	2.66	.75	2.33	.72	2.12	.75	1.99	.81	2.35	.76	188.07*
	가치인식	3.27	.55	3.00	.63	2.72	.73	2.41	.83	2.98	.68	285.20*

\* $p < .05$

### 2. G 연구 분석결과

G 연구 설계를 통해 산출된 분산 및 공분산 성분 추정치는 <표 IV-2>와 같다. <표 IV-2>에 제시된  $p$  분산성분에 대한 행렬에서 대각선에 놓인 원소는 분산 추정치를 나타내며 대각선 상단과 하단의 원소는 각각 측정구인 간 상관관계수 및 공분산 추정치를 나타낸다. 먼저 국어 교과태도에 대한 G 연구 결과를 살펴보면, 성취수준이 높은 집단(우수, 보통)에서 공통적으로 흥미 측정구인에 대한 전집점수 분산( $\sigma_{v_2}^2(p)$ )이 가장 높게 나타났으나 자신감과 가치인식 문항에 대한 분산은 큰 차이를 보이지 않았다. 반면, 성취수준이 낮은 두 집단(기초, 기초미달)에서

는 가치인식 측정구인의 전집점수 분산( $\sigma_{v_3}^2(p)$ )이 가장 높게 나타났다.

<표 VI-2> 성취수준별  $p \times f$  설계의 G 연구 결과

성취 수준	분산 성분	국어			수학			영어		
		자신감	흥미	가치인식	자신감	흥미	가치인식	자신감	흥미	가치인식
우수	p	0.250	0.804	0.642	0.252	0.805	0.671	0.300	0.761	0.609
		0.242	0.364	0.735	0.279	0.477	0.765	0.294	0.497	0.591
		0.161	0.222	0.251	0.161	0.253	0.229	0.164	0.205	0.241
	i	0.021	0.014	0.010	0.034	0.007	0.136	0.019	0.064	0.034
	pi	0.223	0.347	0.231	0.319	0.218	0.400	0.166	0.270	0.259
보통	p	0.237	0.781	0.660	0.247	0.859	0.608	0.370	0.810	0.534
		0.212	0.313	0.739	0.298	0.487	0.732	0.331	0.452	0.548
		0.170	0.219	0.280	0.149	0.252	0.244	0.183	0.208	0.318
	i	0.025	0.018	0.007	0.031	0.013	0.091	0.022	0.044	0.027
	pi	0.224	0.312	0.240	0.399	0.230	0.431	0.205	0.232	0.277
기초	p	0.336	0.855	0.725	0.264	0.905	0.652	0.508	0.899	0.467
		0.316	0.406	0.826	0.339	0.530	0.752	0.458	0.511	0.531
		0.270	0.337	0.410	0.166	0.271	0.246	0.229	0.261	0.473
	i	0.017	0.014	0.010	0.015	0.007	0.010	0.005	0.012	0.017
	pi	0.211	0.252	0.203	0.431	0.198	0.537	0.166	0.187	0.289
기초미달	p	0.464	0.950	0.785	0.259	0.995	0.828	0.611	0.966	0.550
		0.433	0.449	0.828	0.390	0.594	0.959	0.596	0.623	0.637
		0.418	0.434	0.611	0.145	0.255	0.119	0.339	0.396	0.622
	i	0.006	0.006	0.003	0.036	0.000	0.010	0.001	0.004	0.014
	pi	0.234	0.285	0.161	0.536	0.138	0.684	0.180	0.178	0.303

수학과 영어 교과태도의 경우 흥미 문항에 대한 전집점수 분산이 성취수준에 관계없이 가장 높게 나타났다. G연구 결과에서 수학에 대한 가치인식의 분산 추정치를 살펴보면, 기초미달 학생 집단의 경우 피험자 분산이 0.119인 반면, 상호작용(pi)의 분산은 0.684로 나타났으며, 기초 학생 집단에서도 피험자 분산(0.145)에 비해 피험자와 문항 간 상호작용의 분산(0.537)이 상대적으로 크게 나타났다. 또한, 상호작용(pi)의 분산이 성취 수준이 높은 집단일수록 작게 나타난 것을 확인할 수 있다. 세 교과에서 공통적으로 성취수준에 관계없이 전집점수 분산 추정치가 문항에 대한 분산 추정치에 비해 높게 나타나 피험자 간 특성(능력)의 차이가 문항의 특성(난이도)에 의한 차이보다 큼을 알 수 있다. 피험자 특성과 문항 특성 간 상호작용( $\sigma^2(pi)$ )에 대한 분산 추정치는 성취수준이 각기 다른 네 집단에서 대체로 유사한 양상으로 나타났다. 그러나 수학교과 가치인식을 포함한 일부 측정영역에서 특정 성취수준에 대한 피험자와 문항 특

성 간 상호작용( $\sigma^2(pi)$ ) 분산 추정치가 매우 높게 나타나 해당 영역의 측정에 있어 오차 요인이 크게 발생하였으며, 이러한 결과는 이후 살펴 볼 D 연구 분석 결과에서 상대적으로 낮은 신뢰도계수로 반영됨을 확인할 수 있다.

한편, <표 IV-2>에서 각 분산·공분산 행렬의 대각선 상단에 제시된 값을 통해 측정오차를 고려한 상관계수를 살펴보면 성취수준에 관계없이 대체로 .60~.99 범위에 걸쳐 높은 편으로 나타났다. 이는 교과태도의 한 측정구인에 대해 높은 점수를 받은 학생이 다른 측정구인에 대하여도 높은 점수를 받게 되는 상황으로 해석할 수 있다. 특히, 기초미달 집단에서는 흥미와 자신감의 상관이 세 교과에서 각각 .950, .995, .966으로 나타나 성취수준이 낮은 집단일수록 각 교과에 대한 흥미와 자신감의 상관이 높은 것으로 분석되었다.

### 3. D 연구 분석결과

G 연구 설계를 통해 산출된 분산 추정치를 이용하여 계산된 D 연구에서의 측정구인별 분산 및 공분산 추정치와 신뢰도계수는 <표 IV-3>에 제시되어 있다. 국어 교과태도와 관련한 세 개의 측정구인 점수에 대한 일반화가능도 계수는 .801에서 .938로서 모든 성취수준에 걸쳐 대체로 양호하게 나타났다. 이러한 결과는 세 교과 중 국어 교과에 대한 학습 태도 측정이 상대적으로 가장 정확하게 이루어진 것으로 분석된다. 또한, 성취수준별 비교에서는 네 집단 중 기초미달 집단의 가치인식 측정구인에 대한 일반화가능도 계수가 .938로 가장 높게 나타났다. 성취수준을 크게 높고 낮은 두 집단으로 비교할 때, 성취수준이 높은 집단(우수, 보통)에 비해, 성취수준이 낮은 두 집단(기초, 기초미달)에서 전반적으로 국어 교과태도에 대한 측정이 더 정확하게 이루어진 것으로 나타났다. 또한, 국어 교과태도에 대해 성취수준이 가장 높은 ‘우수’ 집단을 제외하고 가치인식에 대한 일반화가능도 계수가 다른 측정구인에 비해 가장 높게 나타난 반면, 흥미와 자신감에 대한 일반화가능도 계수는 대체로 유사하게 나타났다.

수학 교과태도 관련 측정구인별 점수에 대한 일반화가능도 계수의 범위는 .411에서 .945로서 다른 교과에서의 측정과는 달리 성취수준별 신뢰도 및 측정오차의 편차가 상당히 크게 나타났다. 특히, 기초미달 학생 집단의 가치인식 측정구인에 대한 일반화가능도 계수가 .411로 가장 낮게 나타났다. 반면, 같은 집단의 흥미 측정구인에 대한 일반화가능도 계수는 .945로 모든 성취수준에 걸쳐 가장 높게 나타났다. 다른 교과에 대한 측정과는 달리 수학 교과태도 측정에서 성취수준별 그리고 측정구인 간 측정오차 및 신뢰도의 차이가 상대적으로 큰 것으로 분석되었다. 수학 교과 관련 태도와 관련한 세 가지 측정구인에 대해 성취수준과 관계없이 흥미, 자신감, 가치인식 순으로 전집점수의 비율이 높아 결과적으로 신뢰도가 높게 나타난 반면 측정오차는 상대적으로 적게 개입된 것으로 나타났다.

영어 교과태도에 대한 측정 구인별 일반화가능도 계수의 범위는 .789에서 .933으로서 전반적

으로 양호하게 나타났다. 이 중 ‘우수’ 집단을 대상으로 한 영어 교과에 대한 가치인식에 대한 측정에서 일반화가능도 계수가 .789로 가장 낮은 반면, ‘기초미달’ 학생 집단에서 영어 교과 흥미 측정구인에 대한 일반화가능도 계수가 .933으로 이 영역에서의 측정이 가장 신뢰롭게 이루어진 것으로 분석되었다.

<표 VI-3> 성취수준별  $p \times p$  설계의 측정구인 점수에 대한 D 연구 결과

성취 수준	분산성분	국어			수학			영어		
		자신감	흥미	가치인식	자신감	흥미	가치인식	자신감	흥미	가치인식
우수	전 집점수	0.250	0.364	0.251	0.252	0.477	0.229	0.300	0.497	0.241
	상대오차	0.056	0.087	0.058	0.080	0.054	0.100	0.042	0.068	0.065
	절대오차	0.061	0.090	0.060	0.088	0.056	0.134	0.046	0.084	0.073
	일반화가능도계수	0.818	0.807	0.813	0.760	0.898	0.696	0.878	0.880	0.789
	의존도계수	0.804	0.801	0.807	0.741	0.895	0.631	0.866	0.856	0.767
보통	전 집점수	0.237	0.313	0.280	0.247	0.487	0.244	0.370	0.452	0.318
	상대오차	0.056	0.078	0.060	0.100	0.058	0.108	0.051	0.058	0.069
	절대오차	0.062	0.082	0.062	0.107	0.061	0.131	0.057	0.069	0.076
	일반화가능도계수	0.809	0.801	0.824	0.713	0.894	0.693	0.878	0.886	0.821
	의존도계수	0.792	0.791	0.819	0.697	0.889	0.651	0.867	0.868	0.807
기초	전 집점수	0.336	0.406	0.410	0.264	0.530	0.246	0.508	0.511	0.473
	상대오차	0.053	0.063	0.051	0.108	0.049	0.134	0.042	0.047	0.072
	절대오차	0.057	0.067	0.053	0.111	0.051	0.137	0.043	0.050	0.076
	일반화가능도계수	0.864	0.866	0.890	0.710	0.915	0.647	0.924	0.916	0.868
	의존도계수	0.855	0.859	0.885	0.704	0.912	0.643	0.922	0.911	0.861
기초 미달	전 집점수	0.464	0.449	0.611	0.259	0.594	0.119	0.611	0.623	0.622
	상대오차	0.059	0.071	0.040	0.134	0.034	0.171	0.045	0.045	0.076
	절대오차	0.060	0.073	0.041	0.143	0.034	0.174	0.045	0.046	0.079
	일반화가능도계수	0.888	0.863	0.938	0.659	0.945	0.411	0.932	0.933	0.892
	의존도계수	0.885	0.861	0.937	0.644	0.945	0.407	0.931	0.932	0.887

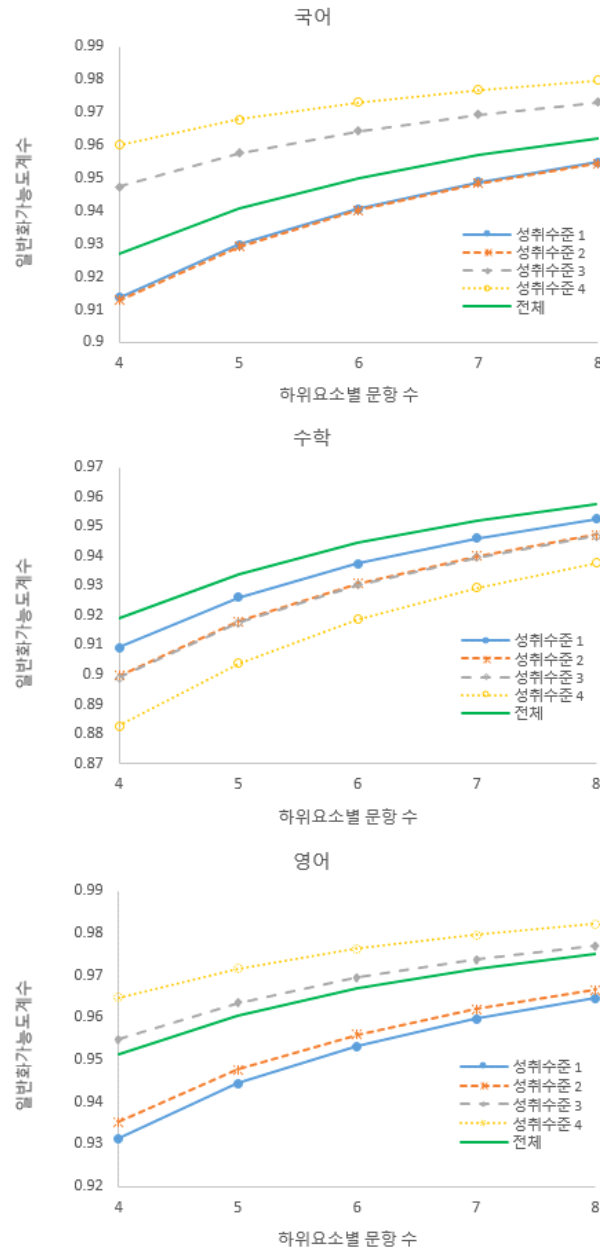
각 교과에 대한 측정구인별 D 연구 결과를 종합하면, 측정오차와 일반화가능도 계수의 패턴에 있어 교과별 차이와 성취수준별 차이가 비교적 두드러지게 나타났다. 국어 교과태도에 대해 성취수준이 가장 높은 ‘우수’ 집단을 제외하고 가치인식에 대한 일반화가능도 계수가 다른 측정구인에 비해 가장 높게 나타난 반면, 흥미와 자신감에 대한 일반화가능도 계수는 대체로 유사하게 나타났다. 이와 대조적으로 수학 교과태도의 경우, 성취수준에 관계없이 흥미에 대한 일반화가능도 계수가 매우 높게 나타난 반면, 자신감과 가치인식은 낮게 나타났다. 특히, 성취수준이 가장 낮은 집단에서 가치인식의 일반화가능도 계수와 의존도계수가 각각 .411과 .407로

다른 집단에 비해 검사 결과의 전반적 신뢰도가 현저히 낮게 나타났다. 이러한 결과는 앞서 G 연구 결과의 해석에서 언급하였듯이 특정 성취수준 집단에 대한 측정에서 피험자와 문항 특성 간 상호작용( $\sigma^2(pi)$ ) 분산 추정치가 매우 높게 산출된 것에 따른 결과로 해석된다. 즉, <표 IV-2>에서 수학교과 가치인식 측정에서 기초미달 학생 집단에 대한  $\sigma^2(pi)$  추정치가 .684로 매우 높게 나타났으며, 이 값은 <표 IV-4>에 제시된 해당 영역에 대한 매우 낮은 일반화가능도 계수 .411을 부분적으로 설명하는 것으로 해석된다.

<표 IV-4>에 제시된 성취수준별 합성점수에 대한 D 연구 결과를 살펴보면, 국어와 영어 교과의 경우 성취수준이 낮은 두 집단에서 합성점수에 대한 일반화가능도 계수가 높게 나타났다. 반면 수학교과와 영어의 경우 성취수준이 높은 집단에서 일반화가능도 계수가 높게 나타나 교과태도에 대한 측정이 해당 집단에서 더욱 정확하고 신뢰롭게 이루어졌음을 알 수 있다. [그림 IV-1]에서는 세 교과태도 영역에서 측정구인별 문항 수를 4개에서 8개 문항으로 증가함에 따른 신뢰도계수와 측정 오차에 미치는 영향력을 분석한 결과를 보여준다.

<표 VI-4> 성취수준별  $\rho^2 \times I^2$ 설계의 합성점수에 대한 D 연구 결과

성취수준	분산성분	교과태도		
		국어	수학	영어
우수	전집점수	0.235	0.261	0.262
	상대오차	0.022	0.026	0.019
	절대오차	0.023	0.031	0.023
	일반화가능도 계수	0.914	0.909	0.931
	의존도계수	0.909	0.894	0.921
보통	전집점수	0.226	0.264	0.287
	상대오차	0.022	0.029	0.020
	절대오차	0.023	0.033	0.022
	일반화가능도 계수	0.913	0.900	0.935
	의존도계수	0.908	0.888	0.928
기초	전집점수	0.333	0.288	0.376
	상대오차	0.019	0.032	0.018
	절대오차	0.020	0.033	0.019
	일반화가능도 계수	0.947	0.899	0.955
	의존도계수	0.944	0.897	0.952
기초미달	전집점수	0.455	0.284	0.502
	상대오차	0.019	0.038	0.018
	절대오차	0.019	0.039	0.019
	일반화가능도 계수	0.960	0.883	0.965
	의존도계수	0.959	0.879	0.964



[그림 IV-1] 측정구인별 문항 수 변화에 따른 일반화가능도 계수의 변화

[그림 IV-1]에서 볼 수 있듯이 성취수준별로 문항 수 증가에 따른 영향력이 각기 다른 양상으로 나타나고 있으며, 세 교과 중 국어에서 그 차이가 비교적 명료함을 알 수 있다. 즉 성취

수준이 낮은 집단에 비해 성취수준이 높은 집단에서 측정구인별 문항 수를 증가함에 따른 신뢰도 향상 효과가 높은 것으로 분석되었다. 문항 수 변화에 따른 신뢰도 향상 효과는 교과 영역과 성취수준에 관계없이 측정 요소별로 4문항에서 5문항으로 증가될 때 그 효과가 가장 큰 것으로 확인되었다. [그림 IV-1]에서 성취수준 집단을 통합한 전체 학생을 대상으로 한 분석 결과를 살펴 보면, 국어와 영어 교과 태도 측정의 결과와는 달리 수학 교과 태도에 대해 일반화가능도계수가 전체 학생 집단에 대해 가장 높게 나타났다. 전체 학생에 대한 자료에서 문항 수 변화에 따른 효과는 성취수준별 분석에서 나타난 결과와 유사한 양상을 보였다.

## V. 결론 및 제언

학업성취도 평가 결과 활용에 앞서 학생들의 다양한 배경변인, 특히 성취 수준에 관계없이 측정이 일관적이며 정확하게 이루어졌는지 점검할 필요가 있다. 이러한 맥락에서 일반화가능도 이론 모형을 적용하여 학업성취도 평가의 정의적 영역에서의 성취수준별 측정의 동등성을 분석하였다. 이를 위해 2015년 중학교 3학년 학생 7,443명을 대상으로 실시한 학업성취도 평가 표집 학생설문 자료 중 교과태도 관련 응답 결과에 대해 다변량 일반화가능도 분석을 실시하였다. 중학교 3학년 표집자료로부터 국어, 수학, 영어 각 교과에 대한 자신감 4문항, 흥미 4문항, 가치인식 4문항을 포함하여 총 12개의 교과태도 관련 변인에 대한 4점 리커트 척도 응답 결과를 분석하였다. 다변량 일반화가능도 분석 모형인  $p^* \times i^o$  설계를 적용하여 G 연구에서는 정의적 특성의 측정 과정에서 발생하는 각 오차요인의 영향력을 분석하였고, 세 교과에서 성취수준별 G 연구 결과를 비교하였다. D 연구에서는 측정구인별로 문항 수를 증가함에 따른 신뢰도와 측정오차에 미치는 영향력을 탐색하였다.

주요 분석 결과는 다음과 같다. 첫째, 학업성취도 평가의 정의적 특성에 대한 측정 도구는 신뢰도와 측정 오차를 기준으로 살펴 볼 때 성취수준의 전 범위에서 걸쳐 교과태도 관련 정의적 성취 특성에 대한 측정이 대체로 양호한 것으로 나타났다. 그러나, 수학 교과에 대한 가치인식 측정에서 기초미달 학생집단의 신뢰도가 .411로 나타나 다른 성취수준 집단과 큰 차이를 보임으로써 측정의 동등성을 담보하기에 다소 부정적인 결과를 보였다. 둘째, 국어, 수학, 영어 교과에서 성취수준별로 각기 다른 양상의 신뢰도와 측정 오차 결과가 나타났다. 즉, 국어 교과 관련 태도에 대해 성취수준이 높은 두 집단에서 공통적으로 흥미 문항에 대한 전집점수 분산이 가장 높게 나타나고 자신감과 가치인식 문항에 대한 분산이 유사하게 나타난 반면, 성취수준이 낮은 두 집단에서는 가치인식의 전집점수 분산이 가장 높게 나타났다. 셋째, 측정오차와 일반화가능도 계수의 패턴에 있어 교과별 차이가 비교적 뚜렷하게 나타나고 있음을 확인하였

다. 국어와 영어 교과태도의 경우 성취수준이 낮은 집단에서 합성점수에 대한 일반화가능도 계수가 높은 반면 수학 교과태도의 경우 성취수준이 높은 집단에서 일반화가능도 계수가 높게 나타나 해당 집단에서 보다 정확한 측정이 이루어진 것으로 분석되었다. 마지막으로, 측정구인별 문항 수를 증가함에 따른 신뢰도와 측정 오차에 미치는 영향력을 분석한 결과, 성취수준별로 그 영향력이 다른 양상으로 나타났으며 세 교과 중 국어에서 그 차이가 비교적 명료하게 나타났다. 즉 성취수준이 낮은 집단에 비해 성취수준이 높은 집단에서 측정구인별 문항 수를 증가함에 따른 신뢰도 향상 효과가 높은 것으로 분석되었다. 또한, 교과태도 관련 정의적 성취 특성 측정을 위한 검사도구의 수정 또는 업데이트가 필요한 경우, 문항 수에 대한 측정 조건 변화를 고려할 때 현재 검사도구에서 측정구인별로 한 문항씩을 추가함으로써 얻게 되는 신뢰도 상승 효과가 가장 크게 나타났다.

이상에서 요약한 연구 결과 중, 본 연구에서 설정한 연구문제와 관련하여 특히 수학 교과에 대한 교과태도 측정 결과에 주목할 필요가 있다. 즉, 수학 교과 가치인식 측정에서 기초미달 학생 집단의 일반화가능도 계수가 .411로 매우 낮고, 기초 학생 집단에서도 자신감과 가치인식의 일반화가능도 계수가 낮게 나타났는데, 이에 대한 주된 원인은 G연구의 분산 성분 중 피험자와 문항 간 상호작용( $\sigma^2(pi)$ ) 분산이 다른 분산 성분에 비해 매우 크게 나타난 것에 기인한 것으로 보인다. 이는 수학 성취도가 낮은 학생들의 경우 수학에 대한 가치인식을 측정하는 영역에서 문항별로 각기 다른 방식으로 응답하거나 비일관적인 동의 정도를 보이고 있음을 의미한다. 이와 같은 현상은 수학 자신감에 있어서도 유사하게 나타났으며, 기초미달 및 기초 학생 집단에서 더욱 두드러지게 나타났다. 가령, 피험자가 개별 문항에 대해 동일한 방식으로 반응하지 않음으로 인하여 피험자별로 각 문항에 대해 동의하는 수준이 다르게 표시될 때, 피험자와 문항 간 상호작용( $\sigma^2(pi)$ ) 분산이 커지게 된다. 상호작용(pi) 분산은 절대오차와 상대오차에 모두 반영되기 때문에, 일반화가능도 계수와 의존도 계수 모두에 영향을 미치게 되어 문항의 신뢰도를 낮추는 결과로 이어지게 된다. 따라서 수학 정의적 특성 중 자신감과 가치인식을 측정하는 개별 문항에 대한 보다 세밀한 내용 검토를 통해 피험자와 문항 간 상호작용을 유발하는 원인을 파악하는 과정이 요구되며, 특히 수학 성취도가 낮은 학생 집단 내 어떠한 피험자 요소가 자신감과 가치인식에 대한 문항에 차별적으로 응답하도록 하였는지에 대한 추가적인 분석을 통해 성취수준이 낮은 학생들의 정의적 특성 향상을 위한 맞춤형 프로그램을 계획하여 제공하는 것이 필요함을 시사한다.

교과태도 관련 정의적 성취 측정의 동등성 분석에 대한 이상의 연구 결과를 종합하면, 일부 측정 영역에서 성취수준과 교과에 따라 오차 요인들의 상대적 영향력이 다르게 나타났으며, 적정 수준의 신뢰도를 확보하기 위한 측정 조건의 변화 효과도 성취수준별로 상이하게 나타났다. 이러한 연구 결과는 다집단 일반화가능도 분석 결과가 피험자 집단별 측정의 동등성 여부를 점검하는 데 유용하게 사용될 수 있음을 의미한다. Li와 Brennan(2007), Powers와



Brennan(2009), Yin(2005) 등 다수의 선행연구에서 보여준 바와 같이, 표준화된 절차를 통해 정교하게 개발된 대규모 검사도구나 할지라도 피험자의 특성과 교과 특성 및 그 외 다양한 검사 요인들에 의해 측정의 정확성 또는 측정오차가 다르게 나타날 수 있으며, 측정오차가 피험자 집단에 따라 다른 방식으로 개입될 경우 측정의 동등성이 위배되는 결과가 초래된다. 따라서, 학업성취도 평가의 정의적 영역에 대한 검사도구 개발 또는 문항 업데이트 과정에서 성취수준 및 교과별 특성과 검사 맥락을 고려한 측정학적 측면에 대한 경험적 검증은 통해 측정의 동등성을 확보하려는 노력이 필요하다고 하겠다. 이러한 실증적 자료 분석 결과를 기반으로 학업성취도 평가의 타당성을 확보함으로써 우리나라 학생들의 정의적 성취 특성을 보다 정확하고 신뢰롭게 파악하는 데 기여할 수 있을 것이다.

본 연구의 제한점에 따른 후속연구를 위한 제언은 다음과 같다. 첫째, 본 연구에서 적용한  $p \times i^\circ$  모형 외에 보다 다양한 검사 요인을 포함하는 다른 모형을 이용하여 분석의 범위를 확장할 수 있다. 가령, 교과태도 관련 정의적 성취 특성의 구성 요소를 국어태도, 수학태도, 영어태도로 고정 한 다변량  $p \times (i^\circ: h^\circ)$  설계를 적용하거나, 피험자가 성취수준 집단에 내재된 자료구조를 고려하여  $(p^\circ: c^\circ) \times i^\circ$  설계를 적용함으로써, 성취수준 국면에 대한 분산과 공분산 성분 추정치, 그리고 교과태도 관련 정의적 성취 특성 요소 간 측정오차를 고려한 상관계수를 비교할 수 있다. 둘째, 본 연구에서는 측정구인(v) 국면을 자신감, 흥미, 가치인식의 세 가지 수준으로만 한정된 고정국면으로 간주하였다. 또한 측정구인별 문항 수도 동일하게 가정하였으나 정의적 성취특성에 대한 보다 심층적인 측정이 요구되는 검사 상황을 고려하여, 측정구인의 수와 측정구인별 문항 수를 보다 자유롭게 갖는 임의국면으로 설정하여 분석할 필요가 있을 것이다. 또한, 본 연구에서는 자신감, 흥미, 가치인식의 세 구인을 고정국면으로 설정하였는데, 정의적 특성을 구성하는 구인은 이 세 구인만이 아니므로 임의국면으로 설정하여 분석하는 것도 가능할 것이다. 마지막으로, 본 연구에서 수행한 D 연구 결과의 활용도를 확장하고 국가수준 학업성취도 평가의 정의적 특성 측정의 정교성 향상에 기여하기 위해 추가적인 분석이 이루어질 필요가 있다. 본 연구에서는 문항 수 변화에 따른 분산성분 및 일반화가능도 계수의 변화를 합성점수 척도 상에서 비교하였으나 후속 연구를 통해 문항 수 변화에 따른 D 연구 결과를 합성점수가 아닌 개별 측정구인별로 실시하여 일정 수준의 신뢰도를 확보하기 위해 각 구인별로 몇 개의 문항이 가장 적합한 수인지에 대한 보다 구체적인 가이드라인을 제공할 필요가 있다. 또한, 본 연구에서는 교과태도 측정 검사도구의 측정구인별 배점 비율을 문항 수에 비례하여 동일하게 설정하였으나, 만약 연구자의 관심이 대규모 표준화 검사 개발에서 측정구인별 가중치를 다르게 설정하는데 있다면 합성점수의 신뢰도에 근거하여 최적의 비율 및 가중치를 결정하는 것이 바람직할 것이다. 즉, 합성점수의 측정구인별 상대 가중치를 다르게 설정하여 그에 따른 일반화가능도 계수의 변화를 확인함으로써 최적의 측정 조건을 탐색하는 후속 연구 수행을 통해 의미 있는 연구결과를 도출할 수 있을 것이다.

## 참 고 문 헌

- 김경희, 김수진 (2010). 수학 및 과학 성취도와 정의적 특성과의 관계에 대한 국제 비교, **교육과정평가연구**, 13(3), 139-208.
- 김성숙, 송미영, 박인용 (2012). 다변량 일반화가능도 이론을 적용한 성취수준 설정에서의 오차 분석과 최적 조건 탐색, **교육평가연구**, 25(4), 679-700.
- 김성연 (2017). 수학적 창의성 태도 검사에서 수학영재와 일반학생의 다집단 일반화가능도 분석, **수학교육논문집**, 31(1), 49-70.
- 김수진, 박지현, 김현경, 진의남, 이명진, 김지영, 안윤경, 서지희 (2012). **수학, 과학 성취도 추이변화 국제비교 연구: TIMSS 2011 결과보고서**. 한국교육과정평가원. 연구보고 RRE 2012-4-3.
- 김양분, 남궁지영, 김정민, 박경호, 임현정 (2014). **한국교육종단연구2013( I ): 초등학생의 교육경험과 교육성취**. 한국교육개발원.
- 김재철. (2002). 학생 배경변인과 수학에 대한 태도변화와의 관계분석: 잠재변인 변화모형의 적용. 서울대학교 박사학위논문.
- 박인용, 이광상, 임해미, 서민희, 김부미, 전경희 (2016). **국가수준 학업성취도 평가의 수학과 정의적 영역 기준 및 지표 산출 방안**. 한국교육과정평가원. 연구보고 RRE 2016-14.
- 이영식, 신상근 (2004). 다변량 일반화가능도 이론에 의한 말하기 시험의 타당도와 신뢰도에 관한 연구. **외국어 교육**, 11(2), 249-265.
- 이진향. (1994). 수학 기피 행동을 야기시키는 부정적인 수학태도의 개선방법에 관한 연구. 이화여자대학교 석사학위논문.
- 이현숙 (2012). 혼합형 검사의 문항 유형별 가중치에 따른 신뢰도 및 다변량 일반화가능도 분석, **교육평가연구**, 25(1), 95-116.
- 전경희, 김성숙 (2017). 정의적 성취특성 측정의 오차요인 및 성취수준별 척도의 동등성 분석. 제9회 국가단위 평가 자료 분석 및 맞춤형 학력진단검사 활용 연구 세미나 발표 자료.
- 최승현, 박상욱, 황혜정 (2014). PISA와 TIMSS 결과에 나타난 우리나라 학생의 정의적 성취 실태 분석 - 수학 교과를 중심으로. **한국학교수학회논문집**, 17(1), 23-43.
- Brennan, R. L. (2001a). *Generalizability Theory*. New York: Springer.
- Brennan, R. L. (2001b). *mGENOVA* [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
- Buckley, J. (2009). *Cross-national response styles in international educational assessments*:

- Evidence from PISA 2006*. Retrieved April 11, 2010, from <https://edsurveys.rti.org/PISA/documents>.
- Lakin, J. M., & Lai, E. R. (2012). Multigroup generalizability analysis of verbal, quantitative, and nonverbal ability tests for culturally and linguistically diverse students. *Educational and Psychological Measurement*, 72(1), 139-158.
- Li, D., & Brennan, R. L. (2007). *A multigroup generalizability analysis of a large-scale reading comprehension test*. CASMA Research Report, 25. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.
- Möttus, R., Allik, J., & Realo, A. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin*, 38(11), 1423-1436.
- OECD (2013). *PISA 2012 Results: Ready to Learn: Students' Engagement, Drive and Self-Beliefs* (Volume III), PISA, OECD Publishing. <http://dx.doi.org/10.1787/9789264201170-en>.
- OECD (2014). *PISA 2012 Results: What Students Know and Can Do - Student Performance in Mathematics, Reading and Science* (Volume I, Revised edition, February 2014), PISA, OECD Publishing. <http://dx.doi.org/10.1787/9789264201118-en>.
- Popham, W. J. (2003). *Test Better, Teacher Better: The Instructional Role of Assessment*. ASCD.
- Powers, S., & Brennan, R. L. (2009). *Multivariate generalizability analyses of mixed format exams*. In Annual Meeting of the National Council on Measurement in Education, San Diego, CA.
- Yin, P. (2005). A multivariate generalizability analysis of the Multistate Bar Examination. *Educational and psychological measurement*, 65(4), 668-686.

## ABSTRACT

### **An Analysis of Measurement Errors and Invariance Properties by Proficiency Level in the Non-cognitive Measures**

**Sungsook Kim**

Senior Research Fellow, Korea Institute for Curriculum and Evaluation

**Kyong Hee Chon**

Assistant Professor, Kangnam University

Before using the results of large-scale student assessments, it is essential to assure that the results are consistent and accurate regardless of various background variables, particularly for the proficiency level. In this context, the purpose of this study was to analyze the measurement equivalence of the non-cognitive ability data across students' groups by the proficiency level based on the generalizability theory approach. This study analyzed student survey data (N=7443) from the 2015 administration of National Assessment of Educational Achievement for ninth graders and applied the multivariate generalizability theory model to the data. In the G study, effects of each error source were investigated, while the focus of the D study was on identifying optimal measurement conditions by varying the number of items within each sub-domain that maximizes the reliability-like indices, such as a generalizability coefficient and an index of dependability. The results of the analysis indicates that the overall measurement was generally reliable over the entire range of the achievement levels; however, the study showed a somewhat negative result in ensuring the measurement equivalence for certain sub-domains of mathematics. In addition, the effect of increasing the number of items to achieve appropriate level of reliability also varied according to achievement level. Based on the results of this study, it was confirmed that a multivariate generalizability analysis approach can be useful for investigating the measurement invariance issues.

*Key Words: Measurement Invariance, Scale Equivalency, Generalizability Theory, Non-cognitive Skills*