교육과정평가연구
The Journal of Curriculum and Evaluation
2013, Vol. 16, No. 2, pp. 149~174

지역독립성 가정의 위배가 IRT 수직척도의 척도변동성에 미치는 영향

박 인 용(한국교육과정평가원 부연구위원)*

≪ 요 약 ≫

문항반응이론(item response theory; IRT)에 기초하여 수직척도를 개발하여 사용할 경우 개발 과정에서 모형, 가정의 충족여부, 자료수집방법 등 많은 요인들이 수직척도 개발에 영향을 끼치기 때문에 그 해석에 유의해야 한다. 이 연구에서는 단위검사로 구성된 검사에서의 자료수집방법 선택에 따라 IRT 방법을 적용하여 수직척도를 개발할 때 지역 독립성 가정의 위배가 수직척도의 특성 중 척도변동성에 미치는 영향을 탐색하고자 모의자료를 통해 공통문항설계방법과 척도검사설계방법에서 수직척도를 개발하였다. 연구결과, 지역독립성 가정이 충족되었을 경우에도 자료수집방법 간 척도변동성이 다르게 산출되는 것을 확인하였다. 또한, 가정이 위배되었을 경우에는 두 자료수집방법 모두에서 척도수축 현상이 나타났으며 그 정도는 공통문항설계방법에서 보다 크게 나타난 것을 확인하였다. 따라서 지역독립성 가정의 위배가 심할 때 공통문항설계방법을 사용하여 자료를 수집할 경우 단위검사모형과 같이 단위검사의 효과를 모형화하여 이를 통제하는 모형을 선택하거나 단위검사를 하나의 측정단위로 하여 이를 다분문항으로 취급해 단위검사의 효과를 없애는 다분문항반응모형 등으로 단위검사로 인한 지역독립성 가정 위배의 영향을 최소화해야 할 것이다.

주제어 : 지역독립성 가정, 단위검사, IRT 수직척도, 척도변동성, 공통문항설계, 척도검사설계

Ⅰ. 서론

학업성취도 평가를 통해 학생의 성장 정도를 적절하게 측정하기 위해서는 학년별 성취도 평

* 제1저자 및 교신저자, iypark@kice.re.kr

가를 통해 산출되는 척도점수들이 서로 비교 가능해야 하며, 각 학년 간에 얻은 척도점수의 변화는 능력수준의 변화로 해석될 수 있어야 한다. 이를 위해서는 학년별 검사의 척도점수를 비교 가능하게 해주는 공통척도의 개발과 각 학년별 성취도 검사의 척도점수들 간 관계 정보를 구하는 작업이 선행되어야 한다(김성훈, 2000; 민경석, 2010; 부재율, 2005; 이규민 외, 2006; 이규민 외, 2010). 이러한 작업을 수직 척도화(vertical scaling)라 한다. 수직 척도화는 같은 구인(construct)을 측정하는 서로 다른 학년의 검사를 공통척도로 변환하는 일련의 과정을 말하며(Dorans, Pommerich, & Holland, 2007; 박인용, 이규민, 강상진, 2012), 이러한 과정을 통해 산출되는 척도를 수직척도라 한다.

수직척도 개발 방법에는 여러 가지가 있으며, 관련 문헌이나 선행연구에서는 수직척도 개발의 모형은 정해져 있지 않으며(Yen, 1986), 가장 적합한 수직척도 개발 방법이나 자료 수집에 대한 합의도 없어(Tong, 2005), 수직척도를 활용하여 학생들의 성장에 대해 해석할 때 수직척도의 개발과정을 충분히 고려하여야 한다고(Briggs & Weeks, 2009) 제시한다. 특히, 문항반응이론(item response theory; IRT)에 기초하여 수직척도를 개발하였을 때는 사용되는 모형, 채점 절차 등 많은 요인들이 수직척도 개발에 영향을 끼치기 때문에 그 해석에 유의해야 한다(Andrews, 1995; Bishop & Omar, 2002; Hendrickson, Kolen, & Tong, 2004; Hendrickson et al., 2005). 이러한 이유로 1970년대부터 80년대 중반까지 IRT 기반의 각 결정사항들에 따라 산출되는 수직척도들의 특성을 탐색하는 연구들이 시행되었다(Kolen, 1981; Marco, Petersen, & Stewart, 1983; Skaggs & Lissitz, 1986). 1980년대 중반에는 IRT 방법을 적용하여 개발한 수직척도의 특성 중에 학년 내 표준편차의 변화인 척도변동성에 주목하였다(Camilli, 1988, 1999; Camilli, Yamamoto, & Wang, 1993; Williams, Pommerich, & Thissen, 1998). 학년이 올라감에 따라 학년 내 표준편차가 줄어드는 현상을 척도수축(scale shrinkage)이라 하는데 많은 학자들이 이러한 현상의 원인에 대해 논의를 하였다.

Camilli(1988)은 학년 내 표준편차가 줄어드는 척도수축이 학년 간 측정 오차가 다르기 때문에 발생하는 것이라 하였고, Camilli 외(1993)은 능력 수준이 매우 높거나 매우 낮은 학생으로 인하여 발생한다고 하였다. Camilli(1999)와 Williams 외(1998)은 모수를 추정할 때 결합최대우도(joint maximum likelihood; JML) 방법 기반인 LOGIST 컴퓨터 프로그램을 사용하게 되면 학년 내 표준편차가 줄어드는 척도수축이 발생한다고 하였으며, 이후 주변최대우도(marginal maximum likelihood; MML) 방법 기반인 BILOG 3 컴퓨터 프로그램(Mislevy & Bock, 1990)을 사용할 때는 발생하지 않았다고 보고하였다. 하지만, Tong(2005)은 Iowa Tests of Basic Skills(ITBS) Form K를 공통문항설계방법에서 IRT 방법을 적용하여 수직척도를 개발하였을 때, MML방법 기반으로 모수추정을 하였지만 지역 독립성 가정의 위배로 인하여 읽기 검사에서 학년에 따른 학년 내 표준편차가 줄어들었다고 보고하였으며,

Tong과 Kolen(2006) 또한 IRT 모형의 일차원성의 가정을 위배하였을 때에는 학년이 올라갈수록 학년 내 표준편차가 줄어드는 척도수축 현상이 나타나지 않지만, 지역 독립성 가정을 위배하였을 때에는 척도수축이 발생한다고 보고하였다. 또한, Kim(2007)도 읽기와 과학검사에서 지역 독립성 가정의 위배로 인하여 학년 내 표준편차가 감소하는 패턴을 보인다고 하였다.

Lee, Brennan과 Frisbie(2000)는 단위검사(testlet)를 검사의 구성, 시행 또는 채점의 과정에서 측정단위로 다루어지는 하나의 문항, 혹은 여러 문항들의 집합이라 정의하며, 이와 관련하여 단위검사의 개념은 IRT 모형의 지역 독립성 가정과 매우 밀접한 관련이 있다고 하였다. 단위검사에 문항이 속해 있는 구조로 검사가 구성이 되어 있을 때 개별적인 문항을 측정의 단위로 사용할 경우 지역 독립성 가정이 위배된다고 보고하고 있다. Tong(2005), Kim(2007) 등은 수직척도를 개발할 때 지역독립성 가정의 위배로 인해 학년 내 표준편차가 감소하는 척도수축 발생의 근거를 실제 데이터를 통해 원인을 추측하고 있다. 따라서, 지역 독립성 가정의 위배 정도와 수직척도의 특성과의 관계에 대한 추가적인 연구가 필요한 실정이다.

수직척도 개발 단계에서 척도의 특성에 영향을 주는 주요한 요인 중 하나는 자료수집방법이다(Andrews, 1995; Kolen & Brennan, 2004; Meng, 2007). 일반적으로 수직척도 개발 시에는 척도검사설계방법과 공통문항설계방법을 사용하는데, 두 방법을 통해 산출되는 수직척도의 특성은 다르게 나타난다(Andrews, 1995; Hendrickson, Kolen, & Tong, 2004). Tong(2005)는 척도검사설계방법으로 개발한 수직척도에서는 학년 내 표준편차의 증감패턴이 일정하지 않으며, 공통문항설계방법으로 개발한 수직척도에서는 학년이 올라갈수록 학년 내 표준편차가 감소하는 패턴을 보인다고 보고하였다. 또한, Tong과 Kolen(2006)은 시뮬레이션 연구를 통해 공통문항설계방법과 척도검사설계방법에서 수직척도를 개발하여 그 특성을 비교하였다. 연구 결과, 공통문항설계방법으로 개발한 수직척도에서는 학년 내 표준편차의 패턴이 감소하였으며, 척도검사설계방법으로 개발한 수직척도에서는 이러한 패턴이 나타나지 않았다. 이와 같이 자료수집방법에 따라 개발되는 수직척도의 특성이 다른 이유는 자료수집방법이 성장의 정의와 밀접하게 연관되어 있으며, 각각의 자료수집방법이 성장의 정의를 반영하고 있기 때문이다. 척도검사를 통해 자료를 수집하는 척도검사설계방법에서는 모든 학년에서 다루는 내용영역에 대한 성장을 반영하고 있으며, 학년 간 공통문항의 자료수집에 중점을 두는 공통문항설계방법에서는 특정 학년에서 다루는 내용에 대한 성장을 반영하고 있다. 이러한 성장의 정의는 검사를 개발하는 시점에서 개발자들이 성장에 대한 철학을 가지고 이를 반영하는 형태로 척도를 개발하기 위해 자료를 수집하는 방법을 선택한다. 따라서, 자료수집방법에 따른 수직척도의 특성을 살펴보아야 할 것이다.

이 연구에서는 단위검사로 구성된 검사에서의 자료수집방법 선택에 따라 IRT 방법을 적용하여 수직척도를 개발할 때 지역 독립성 가정의 위배 정도가 수직척도의 특성 중 척도변동성에 미치는 영향을 탐색하는 것을 목적으로 하고 있으며 보다 세부적인 연구문제는 다음과 같다.

(1) 단위검사로 구성된 검사의 IRT 수직척도에서 지역독립성 가정이 충족될 경우에 공통문항설계방법을 통해 산출된 수직척도의 척도변동성은 어떠한 특성을 보이는가?
(2) 단위검사로 구성된 검사의 IRT 수직척도에서 지역독립성 가정이 충족될 경우에 척도검사설계방법을 통해 산출된 수직척도의 척도변동성은 어떠한 특성을 보이는가?
(3) 단위검사로 구성된 검사의 IRT 수직척도에서 지역독립성 가정의 위배정도에 따라 공통문항설계방법을 통해 산출된 수직척도의 척도변동성은 어떠한 특성을 보이는가?
(4) 단위검사로 구성된 검사의 IRT 수직척도에서 지역독립성 가정의 위배정도에 따라 척도검사설계방법을 통해 산출된 수직척도의 척도변동성은 어떠한 특성을 보이는가?

Ⅱ. 이론적 배경

1. IRT 모형의 가정

IRT에서는 검사에서 재고자 하는 피험자의 잠재적인 특성, 혹은 능력과 문항에 대한 반응과의 관계를 수리적인 함수를 통해 모형화한다. 검사에서 재고자 하는 피험자의 잠재적인 특성, 혹은 능력이 다차원인지, 혹은 단일차원인지에 따라 IRT 모형에 대한 가정은 달라진다. 하지만 다차원 IRT 모형의 경우 추정을 하는데 있어 매우 복잡하며 해석에 있어서도 용이하지 않아 수직척도 개발의 실제 현장에 많이 사용되고 있지 않다. 따라서, 이 연구에서는 검사에서 재고자 하는 피험자의 잠재적 특성, 혹은 능력이 단일차원일 경우만 고려하였으며, 모든 학년의 검사가 이분으로 채점이 되었다고 가정하며 이분 모형일 경우만 고려하였다.

단일차원 이분 IRT 모형은 공통적으로 크게 일차원성과 지역 독립성의 두 가지 기본가정을 바탕으로 하고 있다. IRT 모형의 일차원성 가정은 검사 문항에 대한 피험자의 응답에 반영되어 있는 잠재적 특성, 혹은 능력이 하나라고 가정하는 것이다. 즉, 검사 문항에 의해 오직 하나의 잠재적인 특성, 혹은 능력만 측정된다고 가정하며, 이는 오직 하나의 잠재적 특성, 혹은 능력이 검사 문항에 대한 피험자의 수행수준을 설명한다는 의미이다(Hambleton & Swaminathan, 1985).

지역 독립성 가정은 어떤 능력을 가진 피험자가 특정 문항에 대해 응답할 때 다른 문항에 대한 응답이 영향을 주지 않는다는 가정이다. 이는 피험자가 특정 문항에 답을 맞힐 확률과 다른 문항의 답을 맞힐 확률이 통계적으로 독립인 것을 의미한다. 예를 들면, 세 문항 i, j, k에 대한 피험자의 문항반응을 각각 u_i, u_j, u_k라 하고, 각 문항에 대해 정답으로 응답하면 1, 오답으로 응답하면 0의 값을 갖는다고 가정하면 IRT 모형의 지역 독립성 가정이 충족될 경우 θ_l의 능력

치를 가지고 있는 피험자가 세 문항, i, j, k에 모두 정답으로 응답할 확률은 다음의 식(1)과 같이 표현할 수 있다.

$$P(u_i=1,u_j=1,u_k=1\mid\theta_l)= P(u_i=1\mid\theta_l)P(u_j=1\mid\theta_l)P(u_k=1\mid\theta_l) \quad (1)$$

지역 독립성 가정은 오직 피험자의 능력수준(θ)만이 문항에 정답으로 응답할 확률에 영향을 준다는 의미이며, 이는 일차원성 가정에서 오직 하나의 잠재적 특성, 혹은 능력이 검사 문항에 대한 피험자의 수행수준을 설명한다는 의미와 일맥상통한다. IRT 모형의 일차원성 가정이 충족이 되면 지역 독립성 가정 또한 충족이 되며, IRT 모형의 일차원성과 지역 독립성의 두 가정은 근본적으로 같은 의미를 가진다(Lord, 1980). IRT 모형의 바탕에 있는 일차원성과 지역 독립성의 두 가지 기본가정이 충족이 되면 산출되는 능력모수는 검사 문항의 특성에 의해 영향을 받지 않는 능력모수의 불변성 특성과 문항모수는 피험자의 능력에 의해 변하지 않는 문항모수 불변성의 특성을 가지게 된다. 따라서 원점수나 백분위를 기반으로 수직척도를 개발하는 방법의 한계점을 극복할 수 있는 장점이 있다. 하지만, 일차원성과 지역 독립성 두 가지 가정을 위배할 경우 많은 문제가 발생될 수 있다. 특히, 지역 독립성 가정을 위배하였을 경우 문항모수 추정, 정보함수의 추정, 조건부 측정오차(CSEM) 추정, 신뢰도 추정 등 많은 문제가 발생하게 된다(Lee, 2000; Lee & Frisbie, 1999; Wainer, 1995; Wainer, Bradlow, & Du, 2000; Wang & Wilson, 2000; Wang & Wilson, 2005). 또한, 추정된 문항모수를 기반으로 수행되는 작업인 동등화나 수직척도 개발에도 부정적인 영향을 미치게 된다(Ayala, 2009; Lee, Park, & Jeon, 2009).

2. 단위검사와 수직척도

단위검사는 하나의 문항, 혹은 여러 문항들의 집합이며, 이는 검사의 구성, 시행 또는 채점의 과정에서 측정단위로 다루어진다(Lee et al., 2000). 단위검사로 이루어진 검사의 대표적인 예로는 하나의 문항자료(item material)를 여러 문항이 공유하고 있는 형태로 국어나 영어 검사 등이 있으며, 이러한 검사들은 교육현장에서 일반적으로 널리 사용되고 있다. 교육측정분야에서는 단위검사에 대한 이슈를 1990년대부터 다루어왔다. IRT의 맥락에서 단위검사를 다루는 문제는 단위검사의 측정구조에서부터 시작한다. 단위검사의 측정구조는 하나의 공통된 문항자료에 문항이 내재되어 있는 구조이기 때문에 IRT 일반적인 모형으로의 접근은 단위검사의 측정구조를 설명하지 못하는 한계가 있다. 이러한 측정구조에서는 단위검사에 내재되어 있는 문항 간 독립성을 보장할 수 없기 때문에 문항반응이론의 기본 가정인 지역독립성 가정을 위배하게 되며 이로 인해 모수 추정이나 신뢰도 추정 등에 문제가 발생된다.

Bishop과 Omar(2002)는 단위검사로 이루어진 검사를 통해 지역 독립성 가정의 위배가 수직척도 개발에 미치는 영향에 대해 살펴보고자 하였다. Bishop과 Omar(2002)는 ITBS Form K의 읽기 능력 검사의 3학년에서 8학년에 대한 검사 점수를 이용하여 수직척도를 개발하였다. 지역 독립성 가정을 위배하는 단위검사에 접근하는 방법에 따라 IRT 모형을 이분모형과 다분모형으로 구분하여 총 7개의 IRT 모형을 적용하여 수직척도를 개발하고 개발된 수직척도에서의 능력점수 분포 특성을 살펴보았다. 또한, 이분모형의 경우 동시추정방법과 분리추정방법을 적용하여 수직척도를 개발하였는데, 연구결과 각 모형별로 개발된 수직척도에서의 피험자 점수 분포 표준편차는 학년이 올라갈수록 각 모형별로 패턴의 차이가 나타났다. 3모수 로지스틱 모형, 일반화부분점수모형, 등급반응모형의 경우 표준편차가 줄어드는 패턴을 보였으며, 라쉬모형, 부분점수모형, 명명척도모형의 경우 증감 패턴이 일정하지 않았다. 또한, 동시추정방법과 분리추정방법을 통해 개발된 수직척도에서 모두 학년 내 표준편차가 감소하였다. 특히, 분리추정방법을 통해 개발된 수직척도에서 감소 정도가 심하였다.

Bishop과 Omar(2002)는 이러한 현상을 IRT 모형의 지역 독립성 가정의 위배로 인한 것으로 해석하였다. 하지만, 이 연구는 실제 데이터를 사용하여 지역 독립성 가정의 위배 정도에 따른 수직척도의 특성 차이에 대한 정보는 제공하지 못하는 제한점이 있다. Hendrickson 외(2005)에서도 ITBS 언어와 읽기 검사에 3모수 로지스틱 모형을 적용하여 분리추정방법을 통해 수직척도를 개발하였는데, Bishop과 Omar(2002)와 같은 결과를 산출하였다.

Tong(2005)는 ITBS Form K의 Vocabulary, Math, Reading, Language의 4개 과목의 수직척도를 IRT 방법을 적용하여 개발하였다. 연구결과 공통문항설계방법에서 개발된 수직척도에서 Vocabulary를 제외한 나머지 과목에서 학년이 올라갈수록 학년 내 표준편차가 감소하는 패턴을 보였으며, 이러한 현상의 원인은 IRT 모형의 지역 독립성 가정의 위배로 인한 것이라고 하였다. Ngudgratoke(2006)는 시뮬레이션 연구를 통해 지역 독립성 가정의 위배가 수직척도 개발에 미치는 영향에 대해 살펴보고자 하였다. Ngudgratoke(2006)은 단위검사반응모형(testlet response model; TRM)을 통해 지역 독립성 가정의 위배 정도를 높은 의존성($\sigma^2_{\gamma_{ik(j)}}=1.0$)과 낮은 의존성($\sigma^2_{\gamma_{ik(j)}}=0.5$)으로 구분하여 3개 학년에 대한 데이터를 생성하였다. 생성된 단위검사 데이터를 3모수 로지스틱 모형, 등급반응모형, 단위검사반응모형을 각각 적용하여 공통문항설계방법에서 수직척도를 개발하였다. 연구결과 기존의 단위검사를 다루는 선행연구 결과와는 다소 차이가 있었는데, 3모수 로지스틱 모형이 등급반응모형과 단위검사반응모형보다 정확한 결과를 산출하였다. Ngudgratoke(2006)은 단위검사로 이루어진 검사에서 수직척도를 개발할 때 지역 독립성 가정의 위배 정도에 따라 단위검사를 다루는 모형별로 수직척도의 특성을 살펴보았다는데 의의가 있다. 하지만, 이 연구에서 데이터 생성 시 사용된 높은 의존성의 값이 실제로 지역 독립성 가정의 위배 정도가 큰 값인지 나타내고 있지 않으며 3개 학년의 데이터를 이용해 수직척도를 개발하여 연구결과를 일반화하기에는 무리가 있다. 또한 데이터 생

성모형과 분석모형이 같아 단위검사반응모형으로 분석할 경우 다른 모형을 적용하여 수직척도를 개발할 때 보다 유리하게 작용할 여지가 있다.

Tong과 Kolen(2006)은 IRT 모형의 일차원성과 지역 독립성 두 가지 가정의 위배가 수직척도 개발에 미치는 영향을 살펴보기 위해 시뮬레이션 연구를 수행하였다. ITBS 어휘 검사 결과를 기반으로 하여, 단위검사반응모형과 다차원보상모형(Reckase & McKinley, 1983)을 통해 6개 학년의 지역 독립성 가정을 위반하는 데이터와 다차원 데이터를 생성하였으며 공통문항설계방법과 척도검사설계방법에서 수직척도를 개발하였다. 연구결과 지역 독립성 가정을 위배하였을 경우 공통문항설계방법으로 개발된 수직척도에서 학년 내 표준편차를 과소 추정하였으며, 일차원성 가정을 위배하였을 경우에는 이러한 패턴이 발견되지 않았다. 또한, Kim(2007)도 읽기와 과학검사에서 지역 독립성 가정의 위배로 인하여 학년이 올라갈수록 표준편차가 줄어든다고 보고하고 있다. 이러한 선행연구들은 Tong과 Kolen(2006)을 제외하고는 모두 실제 자료를 이용하여 척도수축현상의 원인을 추측하고 있다. Tong과 Kolen(2006)은 시뮬레이션을 통해 IRT 모형의 기본가정이 위배될 때 개발된 수직척도에 미치는 영향을 살펴보았다는 데 의의가 있다. 하지만, 시뮬레이션 반복과정을 6번으로 제한하였고, 피험자 점수 분포의 표준편차를 모든 학년에 같은 값으로 설정하여 Hoover(1984)가 제시한 학년 내 표준편차가 증가하는 척도개발상황에서의 정보를 제공하지 못하는 제한점이 있어 결과를 일반화하기에는 어려울 수 있다.

Ⅲ. 연구 방법

1. 모의실험

가. 응답자료 생성 모형

이 연구에서는 Nandakumar(1991)가 제시한 다차원 보상모형을 통해 지역독립성 가정의 위배 정도를 모형에 반영하여 모의실험 연구를 시행하였다. Nandakumar(1991)가 제시한 다차원보상모형에서는 피험자의 응답에 두 가지 능력모수가 영향을 주는데, 하나는 모든 문항에 대한 능력모수(general ability;θ_g)이고 다른 하나는 특정한 단위검사에 대한 능력모수(testlet-specific ability;θ_t)이다. 이 모형은 문항 i에 대한 피험자 응답 확률을 다음과 같은 식으로 표현한다.

$$P_i(\theta_{g,}\theta_t) = c_i + \frac{1-c_i}{1+\exp\{-1.7[a_{1i}(\theta_g - b_{1i}) + a_{2i}(\theta_t - b_{2i})]\}} \quad (2)$$

위의 식(2)에서 $P_i(\theta_{g,}\theta_t)$는 검사에서 재고자 하는 능력치가 θ_g이고 특정 단위검사 t에 대한 능력치가 θ_t인 피험자가 특정 단위검사 t에 속한 문항 i에 정답으로 응답할 확률을 의미한다. a_{1i}는 검사에서 재고자하는 능력차원에서의 문항 i의 변별도이며, a_{2i}는 특정 단위검사에 대한 능력차원에서의 문항 i의 변별도이다. b_{1i}는 검사에서 재고자하는 능력차원에서의 문항 i의 난이도이며, b_{2i}는 특정 단위검사에 대한 능력차원에서의 문항 i의 난이도이다. c_i는 문항 i의 추측도이다. 위의 모형을 통해 피험자의 응답 데이터를 생성하기 위해서는 문항 i에 정답으로 응답할 확률에 영향을 주는 모수들을 생성해야 한다.

Nandakumar(1991)에서는 b_{1i}와 b_{2i}의 경우 표준정규분포에서 독립적으로 생성하여 사용하였으며, a_{1i}와 a_{2i}의 경우 다음 식(3)과 같은 분포에서 생성하여 사용하였다.

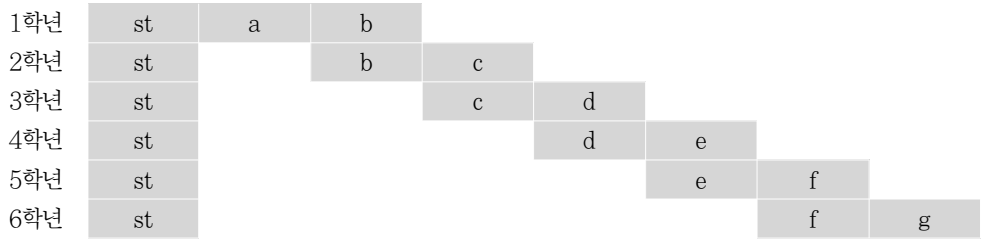
$$\begin{aligned} &a_1 \sim N\{(1-\xi)\mu, (1-\xi)\sigma^2\} \\ &a_2 \sim N(\xi\mu, \xi\sigma^2), \\ &a_1 + a_2 \sim N(\mu, \sigma^2) \end{aligned} \quad (3)$$

위의 식(3)에서 μ와 σ는 정규분포의 평균과 표준편차를 의미하며, ξ값은 피험자의 어떤 특정한 단위검사의 능력수준(θ_t)이 재고자 하는 능력수준 즉, θ_g에 미치는 상대적인 영향력이다. 이는 단위검사에 속한 문항의 의존성 정도이며 지역 독립성 가정의 위배 정도로 해석될 수 있다. 예를 들어, ξ값이 0이면, 피험자가 정답으로 응답할 확률은 θ_g에 의해서만 설명이 되지만, ξ값이 증가되면 θ_t의 값이 증가되어 피험자의 응답에 θ_g보다는 어떤 특정한 단위검사에 대한 능력수준이 미치는 영향력이 커짐을 의미한다.

Lee(2000)에서는 지역독립성 가정이 위배된 실제 데이터를 적절히 반영하는 ξ값을 선택하기 위해 ξ값을 .1부터 .6까지 증가시키며 데이터를 분석하여 Yen의 Q3 지수를 산출하여 비교하였다. 그 결과 ξ값이 .3 근처일 때 지역독립성 가정이 위배된 실제 데이터와 비슷한 값을 산출하여 ξ값을 .2에서 .4사이의 값으로 선택하였다. 또한, Lee와 Park(2008)에서도 ξ값이 .25에서 .35일 때 실제 데이터를 적절히 반영하고 있다고 보고하였다. Lee(2000)에서는 ξ값이 .25일 경우 지역독립성 가정을 '가볍게(mild)' 위배한 정도를 나타내며 .35의 경우 지역독립성 가정을 '심각하게(severe)' 위배한 정도를 나타낸다고 보고한다. 따라서, 이 연구에서도 ξ값을 .00, .25, .30, .35로 설정하여 다차원보상모형을 통해 단위검사에 속한 문항 간 의존성 정도에 따른 피험자 응답 데이터를 생성하였다.

나. 검사구성

이 연구에서는 공통문항설계방법과 척도검사설계방법에서 수직척도를 개발하여 그 특성을 비교하였다. 공통문항설계방법과 척도검사설계방법에서는 각 학년의 학생들만 시행하는 수준검사와 모든 학년의 학생들이 공통적으로 시행하는 척도화 검사를 통해 자료를 수집한다. 이 연구에서는 〔그림 1〕과 같이 척도화 검사와 수준검사를 구성하였다.



〔그림 1〕 학년별 척도화 검사와 수준검사의 구성

st는 척도화 검사를 나타내며, a, b, c, d, e, f, g는 문항집합을 의미한다. 1학년 수준검사의 경우 문항집합 a와 문항집합 b로 구성되며 2학년 수준검사는 문항집합 b와 문항집합 c로 구성된다. 따라서, 문항집합 b는 1학년 수준검사와 2학년 수준검사의 공통문항으로 기능하게 된다. 3학년 수준검사는 문항집합 c와 문항집합 d로 구성되며 이때 문항집합 c는 2학년과 3학년의 공통문항으로 기능한다. 이 연구에서는 〔그림 1〕과 같은 구조로 데이터를 생성하며 척도검사설계방법에서는 모든 학년의 피험자들이 척도화 검사에 응답한 데이터를 통해 수직척도를 개발하고 공통문항설계방법에서는 각 학년별 수준검사만을 이용하여 인접해 있는 학년의 수준검사에 포함되어 있는 공통문항을 통해 수직척도를 개발하였다. 이 연구에서 설정한 각 학년의 수준검사와 척도화 검사를 구성하는 문항집합별 문항구성은 다음과 같다.

〈표 1〉 문항집합별 문항구성

	a	b	c	d	e	f	g	st
단위검사 수	4	5	5	4	4	5	4	7
단위검사에 속한 문항 수	4	4	4	6	6	5	7	4
총 문항 수	16	20	20	24	24	25	28	28

위의 표에서 a, b, c, d, e, f, g, st는 문항집합이며, 이러한 문항집합은 〔그림 1〕과 같이 학년별 척도화 검사와 수준검사를 구성하고 있다. 문항집합 a는 총 4개의 단위검사로 구성되어 있으며, 하나의 단위검사에 4문항으로 구성이 된다. 문항집합 b는 총 5개의 단위검사로 구성되

어 있으며, 하나의 단위검사에 4문항이 포함되어 있다. 1학년 수준검사의 경우 문항집합 a와 문항집합 b로 구성이 된다. 따라서, 1학년 수준검사는 총 9개의 단위검사로 구성이 되어 있으며, 하나의 단위검사에 4문항씩 포함되어 총 36문항으로 구성되어 있다. 2학년 수준검사는 문항집합 b와 문항집합 c로 구성되어 있으며, 3학년 수준검사는 문항집합 c와 문항집합 d로 구성되어 있다. 이러한 문항집합의 구성과 6개 학년에 대한 구성은 ITBS Form K의 Reading Comprehension 검사 구조를 기반으로 하여 단위검사에 속해 있는 문항의 균형을 고려하였으며, Kolen과 Brennan (2004)에서 제시한 척도검사설계방법의 구성을 고려하여 설정하였다.

다. 문항 및 피험자 모수

이 연구에서는 지역독립성 가정의 위배정도에 대한 정보가 포함되어 있는 다차원 보상모형을 통해 데이터를 생성하였다. 따라서, 다차원보상모형에 포함되어 있는 문항모수들을 명세화할 필요가 있다. 각 문항집합별 문항모수들의 분포를 ITBS Form K의 Vocabulary를 기반으로 설정하였으며 다음의 〈표 2〉에 제시하였다.

〈표 2〉 문항집합에 속한 문항모수 분포

정규분포		a	b	c	d	e	f	g	st
a_1 a_2	평균	0.95	1.07	0.94	1.16	1.14	1.13	1.19	1.28
	표준편차	0.13	0.31	0.26	0.33	0.33	0.29	0.41	0.42
b_1 b_2	평균	-0.90	0.08	0.59	1.42	1.37	2.02	2.13	1.00
	표준편차	0.74	0.67	0.62	0.47	0.64	0.54	0.60	1.19
c	평균	0.15	0.17	0.18	0.17	0.18	0.17	0.19	0.20
	표준편차	0.04	0.06	0.05	0.04	0.04	0.02	0.06	0.05

모의실험 자료를 생성하기 위해 학년별 피험자 능력점수 분포를 〈표 3〉과 같이 상정하였으며, 이는 학년이 올라갈수록 평균 능력수준이 증가하는 형태를 가지고, 표준편차 또한 증가하여 높은 능력수준의 학생이 낮은 능력수준의 학생보다 성장정도가 크게 나타나는 것을 반영하고 있다.

〈표 3〉 학년별 피험자 능력점수 분포

	1학년	2학년	3학년	4학년	5학년	6학년
평균	0	0.4	0.8	1.2	1.6	2
표준편차	1	1.1	1.2	1.3	1.4	1.5

라. 응답자료 생성 및 분석 절차

이 연구에서는 〔그림 1〕과 같은 문항집합을 통해 척도화 검사와 수준검사를 구성하였으며 각 문항집합별 단위검사는 〈표 1〉과 같이 구성하였다. 또한 문항집합별 문항모수의 분포를 〈표 2〉와 같이 설정하고 〈표 3〉과 같은 학년별 피험자 능력점수 분포에서 각 학년별 1,500명에 대한 능력모수를 생성하였다. 산출된 문항모수, 능력모수, 지역 독립성 가정의 위배 정도(ξ = .00, .25, .30, .35)를 식(2)와 식(3)에 적용하여 피험자가 문항에 정답으로 응답할 확률을 통해 응답자료를 생성하였다. 각 학년별로 수준검사와 척도화 검사에 대한 응답자료를 생성하였으며, 이러한 절차를 100번 반복하여 생성한 피험자 응답자료를 자료수집방법에 따라 수직척도를 개발하여 그 특성을 비교하였다.

이 연구에서는 공통문항설계방법의 경우 〔그림 1〕에서 척도화 검사의 응답자료를 제외하고 수준검사만을 이용하여 각 학년별 문항모수를 MMLE 방법 기반인 BILOG-MG(Zimowski, Muraki, Mislevy & Bock, 2003)프로그램을 통해 분리 추정하였으며, 인접학년 간 공통문항을 통해 Stocking-Lord 방법을 통해 1학년을 기준척도로 하여 공통척도를 개발하였다. 척도검사설계방법의 경우 〔그림 1〕의 모든 응답자료를 사용하여 수직척도를 개발하였다. 먼저 모든 학년의 피험자가 척도화 검사에 응답한 자료만을 이용하여 BILOG-MG(Zimowski et al., 2003)를 통해 1학년을 기본척도로 하여 추정(calibration)을 시행한 후, 각 학년의 능력점수 분포의 평균과 표준편차를 활용하여 각 학년의 수준검사에 대한 모수추정을 시행하였다. 이와 같은 절차로 산출되는 피험자 능력점수 분포의 표준편차를 통해 척도변동성에 대한 양호도를 산출하였다.

2. 평가준거

이 연구에서는 자료수집방법별로 개발된 수직척도에서의 척도변동성에 대한 양호도를 살펴보기 위해 세 가지 지표를 사용하였다. RMSE(root mean squared error)는 추정치와 실제 모수의 차이를 나타내며, BIAS는 추정치의 평균과 실제 모수와의 차이를 나타낸다. SEE (standard error of estimates)는 추정치와 추정치 평균의 차이를 나타내며, 세 지표는 식 (4)로 산출된다.

$$
\begin{aligned}
RMSE_g &= \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\tau}_{gr}-\tau_g\right)^2} & RMSE &= \frac{1}{G}\sum_{g=1}^{G} RMSE_g \\
BIAS_g &= \left(\bar{\hat{\tau}}_g-\tau_g\right) & BIAS &= \frac{1}{G}\sum_{g=1}^{G}|BIAS_g| \\
SEE_g &= \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\tau}_{gr}-\bar{\hat{\tau}}_g\right)^2} & SEE &= \frac{1}{G}\sum_{g=1}^{G} SEE_g
\end{aligned}
\qquad (4)
$$

g는 학년을 뜻하며, r은 반복을 의미한다. $\hat{\tau}$는 척도변동성의 추정치를 의미하며 τ는 실제 모수(true value)를 의미한다. 자료수집방법별로 개발된 수직척도에서의 각 학년별 피험자 능력점수 분포의 표준편차가 $\hat{\tau}_g$가 되며 〈표 1〉에 명시된 각 학년별 표준편차가 τ_g가 된다. $RMSE_g$, $BIAS_g$, SEE_g는 각 학년별로 산출이 되며, 모든 학년에 대한 $RMSE_g$, $BIAS_g$, SEE_g의 평균은 종합적인 지표로 사용될 수 있다. $RMSE_g$, $BIAS_g$, SEE_g는 학년별로 추정치의 양호도, 과소 혹은 과대 추정 정도, 안정성에 대한 정보를 주고, RMSE, BIAS, SEE는 모든 학년에 대한 평균적인 양호도, 정확성, 안정성에 대한 정보를 제공한다.

RMSE는 각각의 방법으로 개발된 수직척도에서의 척도변동성에 대한 양호도 정도를 나타내는 지표이며, RMSE가 작을수록 양호도가 좋은 것을 의미한다. BIAS는 각각의 방법에 따라 산출되는 척도변동성이 모든 학년에 대해 평균적으로 실제값과 얼마나 차이가 있는지를 보여주며, 수직척도를 개발하는 방법에서의 척도변동성에 대한 정확성을 나타낸다. BIAS가 작을수록 체계적 오차가 작은 것을 의미하며 수직척도를 개발하는 방법이 척도변동성의 추정 측면에서 정확하다는 것을 의미한다. SEE는 추정치의 안정성 정도에 대한 정보를 준다. SEE가 작은 방법이 개발된 수직척도에서 안정적인 척도변동성 추정치를 산출하는 것을 나타낸다.

Ⅳ. 연구 결과

1. 공통문항설계방법을 통한 수직척도에서의 척도변동성 양호도

지역독립성 가정의 위배정도에 따라 생성된 인접학년 간의 공통문항에 대한 응답데이터를 통해 산출된 수직척도에서의 척도변동성은 〈표 4〉에 제시하였다.

〈표 4〉 공통문항설계방법에서 ξ값에 따른 학년별 표준편차

ξ값	1학년	2학년	3학년	4학년	5학년	6학년
.00	1.000	1.100	1.206	1.324	1.399	1.500
.25	1.000	1.090	1.181	1.286	1.373	1.456
.30	1.000	1.087	1.186	1.286	1.375	1.457
.35	1.000	1.097	1.195	1.294	1.388	1.462

공통문항설계방법을 통해 지역독립성 가정의 위배정도에 따른 데이터를 기반으로 수직척도를 개발하였을 때 학년이 올라감에 따라 표준편차가 증가하는 패턴을 보였으며, 이러한 결과는 척

도변동성의 실제모수가 증가하는 패턴을 가지고 있기 때문에 나타나는 당연한 결과이다. 산출된 척도변동성의 학년별 RMSE, BIAS, SEE 및 평균 양호도 정보는 〈표 5〉에 제시하였다.

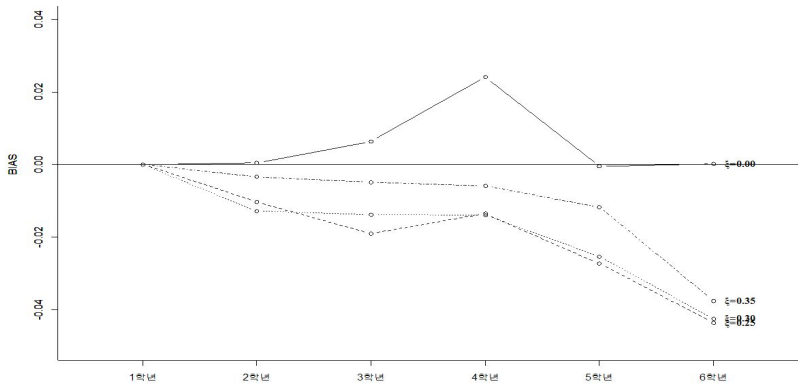
〈표 5〉 공통문항설계방법에서의 ξ값에 따른 척도변동성의 RMSE, BIAS, SEE

평가준거	ξ값	1학년	2학년	3학년	4학년	5학년	6학년	평균
BIAS	.00	0.000	0.000	0.006	0.024	-0.001	0.000	0.005
	.25	0.000	-0.010	-0.019	-0.014	-0.027	-0.044	0.019
	.30	0.000	-0.013	-0.014	-0.014	-0.025	-0.043	0.018
	.35	0.000	-0.003	-0.005	-0.006	-0.012	-0.038	0.011
SEE	.00	0.000	0.025	0.042	0.056	0.065	0.080	0.045
	.25	0.000	0.031	0.049	0.063	0.078	0.081	0.050
	.30	0.000	0.031	0.053	0.073	0.086	0.100	0.057
	.35	0.000	0.041	0.060	0.077	0.091	0.104	0.062
RMSE	.00	0.000	0.025	0.042	0.061	0.065	0.080	0.045
	.25	0.000	0.033	0.052	0.065	0.082	0.092	0.054
	.30	0.000	0.033	0.055	0.074	0.089	0.109	0.060
	.35	0.000	0.041	0.061	0.078	0.092	0.111	0.063

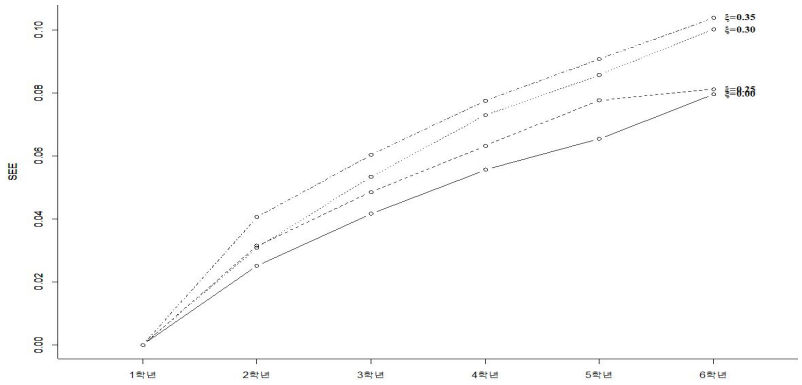
척도변동성 추정의 정확성 정도를 보여주는 BIAS는 모든 학년에 대해 평균적으로 약 0.005 ~ 0.019로 산출되었으며, ξ값에 따라 학년별 패턴은 다르게 나타났다. 척도변동성 추정의 안정성을 보여주는 SEE는 모든 학년에 대해 평균적으로 ξ값에 따라 약 0.045에서 0.062로 나타났으며, ξ값에 따라 그 값이 점차 커져 지역독립성 가정의 위배 정도가 클수록 척도변동성 추정의 모든 학년에 대한 평균적인 안정성은 떨어지는 것을 보여주고 있다. 척도변동성 추정의 전반적인 양호도를 보여주는 RMSE를 살펴보면, 모든 학년에 대해 평균적으로 약 0.045 ~ 0.063으로 나타났으며, 추정의 안정성에 의한 영향으로 ξ값에 따라 점차 그 값이 커져 전반적인 양호도가 지역독립성 가정의 위배정도에 따라 감소하는 것을 볼 수 있다. ξ값별 척도변동성의 BIAS, SEE 패턴은 〔그림 2〕와 〔그림 3〕에 제시하였다.

〔그림 2〕의 X축은 학년을 나타내며, Y축은 BIAS를 보여주고 있다. 공통문항설계방법을 통해 자료를 수집하고 척도를 개발하였을 때의 ξ값별 학년에 따른 BIAS 패턴을 살펴보면, 지역독립성 가정을 충족하였을 경우(ξ=0.00) BIAS는 학년별로 약 -0.001 ~ 0.024로 산출되었고 평균 BIAS도 약 0.005로 매우 작게 나타나 학년별로 거의 정확하거나 약간의 과대 추정이 있는 것을 볼 수 있었다. 반면에 지역독립성 가정이 위배되었을 경우(ξ=0.25, 0.30 ,0.35)에는 BIAS가 약 -0.003 ~ -0.044로 모두 과소 추정하는 것을 볼 수 있었으며, 그 정도는 학년별, 지역독립성 가정의 위배 정도별로 다르게 나타났다. 이러한 점은 공통문항설계방법을 통해

산출된 척도변동성이 지역독립성 가정의 위배에 따라 과소 추정되며, 이는 척도수축으로 연결되는 것을 보여준다. 또한, ξ값이 0.35일 때 0.25, 0.30일 때보다 오히려 양호도가 BIAS가 작게 나타나는 역전현상이 나타났는데, 이러한 결과는 Tong과 Kolen (2007)의 결과에서도 볼 수 있었다. 이러한 점으로 보아 척도수축은 지역독립성 가정의 위배 정도보다는 위배여부에 보다 민감하게 영향을 받는 것으로 판단된다.



〔그림 2〕 공통문항설계방법에서의 ξ값에 따른 척도변동성의 BIAS



〔그림 3〕 공통문항설계방법에서의 ξ값에 따른 척도변동성의 SEE

〔그림 3〕은 ξ값별 학년에 따른 척도변동성 추정의 SEE의 값을 보여주고 있는데, 척도변동성 추정의 안정성을 보여주는 SEE의 경우 모든 ξ값에서 학년이 높아질수록 SEE가 올라가는 패턴을 보이고 있다. 지역독립성 가정을 충족하였을 경우(ξ=0.00) SEE는 학년별로 약 0.025 ~ 0.08로 산출되었고 평균 SEE는 약 0.045로 나타났다. 반면에 지역독립성 가정이 위배되었을 경우(ξ=0.25, 0.30 ,0.35)에는 SEE가 약 0.031 ~ 0.104로 산출되었으며, 평균 SEE

도 0.054 ~ 0.063으로 나타났다. 이러한 점은 1학년을 기준척도로 하여 공통문항설계방법을 통해 산출된 척도변동성 추정치의 안정성 측면의 오차가 학년에 따라 누적되어 나타나는 현상으로 판단된다. 또한 지역독립성 가정의 위배 정도의 영향으로 인해 추정치의 안정성이 보다 떨어지고 있음을 보여주고 있다. 척도변동성 추정치의 전반적인 양호도를 보여주는 지표인 RMSE는 SEE와 비슷한 패턴을 보이고 있으며, 지역독립성 가정을 충족하였을 경우보다 위배하였을 경우에 모든 학년에서 전반적인 양호도가 낮게 나타났다.

2. 척도검사설계방법을 통한 수직척도에서의 척도변동성 양호도

척도검사설계방법에서는 척도검사를 통해 수직척도를 개발하며, 이를 반영하여 지역독립성 가정의 위배정도에 따라 생성된 자료를 통해 산출된 수직척도에서의 척도변동성은 〈표 6〉에 제시하였다.

〈표 6〉 척도검사설계방법에서 ξ값에 따른 학년별 표준편차

ξ값	1학년	2학년	3학년	4학년	5학년	6학년
.00	1.000	1.104	1.208	1.308	1.403	1.483
.25	1.000	1.104	1.206	1.316	1.425	1.528
.30	1.000	1.103	1.202	1.314	1.412	1.507
.35	1.000	1.100	1.193	1.286	1.388	1.482

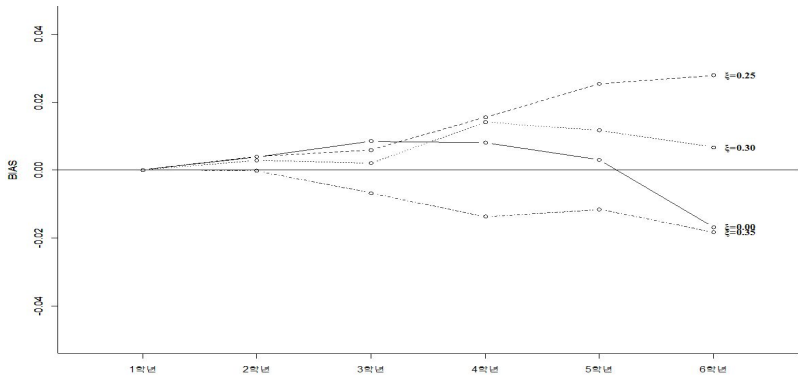
척도검사설계방법을 통해 지역독립성 가정의 위배정도에 따른 자료를 기반으로 수직척도를 개발하였을 때 학년이 올라감에 따라 표준편차가 증가하는 패턴을 보였으며, 이러한 결과는 공통문항설계방법을 통해 산출된 결과와 동일하게 척도변동성의 실제모수가 증가하는 패턴을 가지고 있기 때문에 나타나는 당연한 결과이다. 척도검사설계방법을 통해 산출된 척도변동성의 학년별 RMSE, BIAS, SEE 및 평균 양호도 정보는 〈표 7〉에 제시하였다.

〈표 7〉 척도검사설계방법에서의 ξ값에 따른 척도변동성의 RMSE, BIAS, SEE

평가준거	ξ값	1학년	2학년	3학년	4학년	5학년	6학년	평균
BIAS	.00	0.000	0.004	0.008	0.008	0.003	-0.017	0.007
	.25	0.000	0.004	0.006	0.016	0.025	0.028	0.013
	.30	0.000	0.003	0.002	0.014	0.012	0.007	0.006
	.35	0.000	0.000	-0.007	-0.014	-0.012	-0.018	0.008

평가준거	ξ값	1학년	2학년	3학년	4학년	5학년	6학년	평균
SEE	.00	0.000	0.024	0.025	0.028	0.035	0.037	0.025
	.25	0.000	0.040	0.039	0.041	0.048	0.052	0.037
	.30	0.000	0.037	0.033	0.039	0.043	0.045	0.033
	.35	0.000	0.036	0.038	0.042	0.045	0.049	0.035
RMSE	.00	0.000	0.024	0.027	0.029	0.035	0.041	0.026
	.25	0.000	0.040	0.040	0.043	0.055	0.059	0.040
	.30	0.000	0.037	0.033	0.041	0.044	0.045	0.034
	.35	0.000	0.036	0.038	0.044	0.046	0.052	0.036

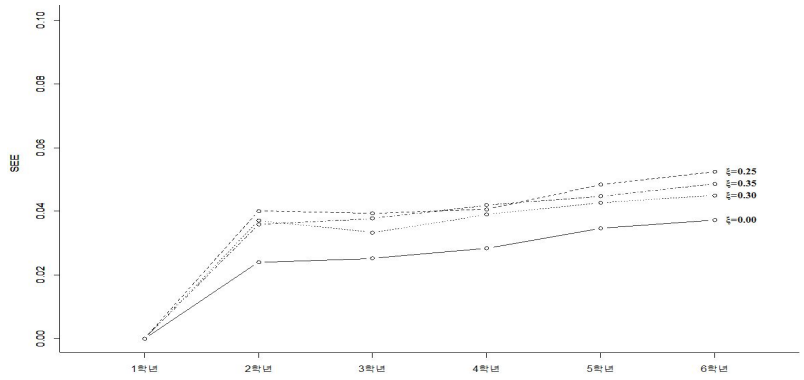
척도검사설계방법에서 산출된 척도변동성의 BIAS는 모든 학년에 대해 평균적으로 약 0.006 ~ 0.013으로 산출되어 공통설계방법에서의 척도변동성에 대한 평균 BIAS보다 작게 나타난 것을 볼 수 있다. 척도변동성 추정의 안정성 지표인 SEE는 모든 학년에 대해 평균적으로 ξ값에 따라 약 0.025에서 0.037로 나타나, 공통설계방법에서의 척도변동성에 대한 평균 SEE 보다 작게 나타난 것을 볼 수 있다. 또한, 지역독립성 가정이 충족되었을 경우 위배되었을 경우에 비해 상대적으로 평균 SEE가 낮게 나타난 것을 볼 수 있다. 척도변동성 추정의 RMSE를 살펴보면, 모든 학년에 대해 평균적으로 약 0.026 ~ 0.040으로 나타나, BIAS, SEE와 동일하게 공통설계방법에서의 척도변동성 추정치보다 양호도가 좋은 것을 나타났다. ξ값별 척도변동성의 BIAS, SEE 패턴은 〔그림 4〕와 〔그림 5〕에 제시하였다.



〔그림 4〕 척도검사설계방법에서의 ξ값에 따른 척도변동성의 BIAS

〔그림 4〕는 척도검사설계방법에서의 척도변동성 추정에 대한 BIAS를 보여주고 있는데, X축은 학년을 나타내며, Y축은 BIAS를 나타낸다. 척도검사설계방법을 통해 자료를 수집하고 척도

를 개발하였을 때의 ξ값별 학년에 따른 BIAS 패턴을 살펴보면, 지역독립성 가정을 충족하였을 경우(ξ=0.00) BIAS는 학년별로 약 -0.017 ~ 0.008로 산출되었고 평균 BIAS도 약 0.007로 작게 나타나 학년별로 거의 정확하나 약간의 과대 및 과소 추정이 있는 것을 볼 수 있었다. 지역독립성 가정이 위배되었을 경우(ξ=0.25, 0.30 ,0.35)에는 모든 학년에 평균적인 BIAS는 가정이 충족되었을 경우와 큰 차이를 보이지는 않았지만 학년별 BIAS가 약 -0.018 ~ 0.028로 나타나 학년에 따른 과소 및 과대 추정 정도는 그 패턴에 있어 차이를 보이고 있었다. 이러한 점은 척도검사설계방법에서의 척도변동성 추정은 공통문항설계방법에서보다 모든 학년에 평균적으로 정확하고, 안정적으로 수행되는 것을 보여준다.



〔그림 5〕 척도검사설계방법에서의 ξ값에 따른 척도변동성의 SEE

척도검사설계방법에서의 ξ값별 학년에 따른 척도변동성 추정의 SEE의 경우 모든 ξ값에서 1학년에서 2학년으로 올라갈 때 가장 크게 값이 증가하고 있으며, 이후 학년으로 갈수록 증가폭은 상대적으로 낮게 나타나고 있다. 이러한 점은 척도검사설계방법에서 학년에 따른 안정성 측면의 오차가 누적되는 경향이 공통설계방법에서보다 적은 것을 보여준다. 척도검사설계방법에서 지역독립성 가정을 충족하였을 경우(ξ=0.00) SEE는 학년별로 약 0.024 ~ 0.037로 산출되었고, 평균 SEE는 약 0.025로 나타났다. 반면에 지역독립성 가정이 위배되었을 경우(ξ=0.25, 0.30 ,0.35)에는 SEE가 약 0.033 ~ 0.052로 산출되었으며, 평균 SEE도 0.033 ~ 0.037로 나타났다. 지역독립성 가정 위배의 영향으로 인해 추정치의 안정성이 보다 떨어지고 있음을 보여주고 있으며, 그 영향은 공통문항설계방법보다 적은 것을 보여준다. 또한, 그림으로는 제시하지 않았지만 척도변동성 추정치의 전반적인 양호도를 보여주는 지표인 RMSE는 SEE와 비슷한 패턴을 보이고 있으며, 지역독립성 가정을 충족하였을 경우보다 위배하였을 경우에 RMSE가 높게 나타나 전반적인 양호도가 모든 학년에서 좋은 것을 볼 수 있다.

Ⅴ. 결론 및 논의

IRT 방법을 적용하여 수직척도를 개발하는 과정은 매우 복잡하며 사용되는 모형과 가정의 충족 여부, 자료수집방법 등이 영향을 준다. 이 연구는 IRT 모형의 지역독립성 가정을 위배하였을 때 산출되는 수직척도의 척도변동성에 초점을 두고 지역독립성 가정의 영향에 대해 살펴보았다. 공통문항설계방법과 척도검사설계방법에서 IRT 지역독립성 가정의 위배가 산출되는 척도변동성에 어떠한 영향을 주는지 확인하고자 하였으며, 연구결과를 중심으로 결론과 논의를 제시하면 다음과 같다.

첫째, 지역독립성 가정이 충족되었을 경우 척도변동성 오차 패턴은 자료수집방법 간 다르게 산출되었다. 또한, 자료수집방법별 학년에 따른 척도변동성의 양호도가 불규칙한 형태를 보이고 있었다. 이는 지역독립성 가정이 충족되었을 경우에도 수직척도를 개발하기 위한 자료수집방법으로 인해 수직척도에서의 피험자 능력점수 분포 특성이 다르게 나타나는 것을 보여주고 있다. 따라서 IRT 수직척도가 지역독립성 가정이 충족되었다 하더라도 많은 요인에 의해 영향을 받기 때문에 개발 당시 사용된 동일한 방법을 지속적으로 적용하여 유지해야 할 것이다. 또한, 수직척도 점수의 활용에 있어서도 수직척도 점수의 오차에 대한 정보를 수요자들에게 제공해야 하며, 동일학생이나 학교가 여러 해를 거치며 성장하는 부분에 대한 파악이나, 다년간에 걸친 학업성취에 대한 학교효과 등의 해석에 수직척도 개발 방법의 특성이 고려될 필요가 있다.

둘째, 지역독립성 가정의 위배로 인해 척도수축 현상이 발생하고 있었다. Kim(2007), Tong(2005), Tong과 Kolen(2006)은 공통문항설계방법에서 수직척도를 개발할 경우 지역독립성 가정의 위배로 인해 척도수축이 발생된다고 보고하였다. 이 연구에서도 이러한 결과를 확인할 수 있었으며 지역독립성 가정의 위배가 심해질수록 척도수축의 정도가 커지고 있었다. 또한, 미미하지만 척도검사설계방법에서도 척도수축 현상이 나타났다. 이는 지역독립성 가정이 심할 경우 공통문항설계방법과 척도검사설계방법에서 모두 척도수축이 발생하는 것을 나타내고 있으며 특히, 공통문항설계방법이 지역독립성 가정의 위배에 영향을 많이 받는 것을 보여주고 있다. Hoover(1984)는 학년이 올라가면서 능력수준이 낮은 학생의 수가 능력수준이 높은 수준의 학생 수보다 빨리 증가하기 때문에 학년 내 표준편차가 줄어드는 척도는 문제가 있다고 지적하였다. 이러한 관점에서 보았을 때 실제 척도변동성의 모수가 증가할 경우 지역독립성 가정의 위배가 심할 때 모든 방법에서 척도수축이 발견되는 것은 시사하는 바가 크다. 이러한 결과는 IRT 수직척도가 지역독립성 가정의 위배로 인하여 실제 학생들의 성장을 반영하지 못할 수 있음을 보여주고 있다. 따라서, 문항을 개발하고 검사를 구성하는 시점부터 지역독립성 가정의 위배를 통제할 필요가 있으며, 검사 개발 시 이를 통제하지 못하더라도 통계적인 방법을 통해 단위검사의 효과를 최소화할 필요가 있다. Bishop과 Omar(2002)는 단위검사의 효과를 통제

하기 위해 단위검사를 다분문항으로 처리하였으며, Wainer 외 (2000)과 Wang 외 (2005)는 단위검사의 효과를 통제하는 모형을 제시하였다. 수직척도 개발 시에도 지역독립성 가정의 위배가 심할 때 공통문항설계방법을 사용하여 자료를 수집할 경우 이와 같이 단위검사모형과 같이 단위검사의 효과를 모형화 하여 이를 통제하는 모형을 선택하거나 다분문항반응모형 등의 적용을 통해 단위검사로 인한 지역독립성 가정 위배의 효과를 통제하는 것이 필요하다. 또한, 단위검사의 효과를 통제하는 모형을 통해 수직척도를 개발하기에 앞서 단위검사를 다루는 모형들을 통한 수직척도 개발의 방법론적인 타당성에 대한 연구가 이루어져야 할 것이다.

셋째, 공통문항설계방법보다 지역독립성 가정의 위배에 상대적으로 척도검사설계방법이 영향을 덜 받았다. 이러한 점은 공통문항의 연계 강도의 차이에서 볼 수 있는데, 연계 강도의 차이는 두 자료수집방법에서 수직척도를 개발할 때 사용하는 검사에서 볼 수 있다. 각 학년 간 척도를 공통척도로 변환하기 위해서 척도검사설계방법에서는 척도화 검사가 각 학년 간 공통문항의 기능을 하며 공통문항설계방법에서는 각 인접해 있는 학년의 수준검사에 포함된 공통문항을 통해 각 학년 간 척도를 공통척도로 변환한다. 이 연구에서는 척도화 검사를 구성하는 문항의 수가 각 인접해 있는 학년 간 공통문항의 수보다 크도록 구성하였으며, 이는 일반적으로 실제 검사기관에서 수직척도를 개발할 때 사용하는 구성이기도 하다. 따라서, 보다 많은 수의 문항을 통해 수직척도를 개발하는 척도검사설계방법이 공통문항설계방법보다 안정적인 결과를 산출하는 것으로 보인다.

척도검사설계방법을 통해 수직척도를 개발하는 것이 이론적으로 적절할 수 있지만 척도검사설계방법을 통해 수직척도를 개발하여 실제 적용하는 데에는 어려움이 있을 수 있다. 먼저 척도화 검사를 추가적으로 개발하고 시행해야 하며, 모든 학년의 내용을 반영해야 하기 때문에 검사가 길고 학년이 낮은 학생들의 수준에 맞지 않는 문항들도 포함이 되어있다. 검사에 포함된 문항이 학생의 수준과 맞지 않을 경우 학생들이 이러한 문항에 대해 대부분 추측하여 문항에 응답하게 될 수 있다. 추측에 기초하여 문항에 응답할 경우 문항반응패턴의 이상 범위가 산출될 가능성이 높고, 문항반응패턴이 이상한 경우 문항모수 추정 시 수렴되지 않는 경우가 발생하며 실제 능력점수를 과대 추정할 가능성이 있다. 또한, 여러 학년에 걸친 척도화 검사에 대한 각 학년 학생들의 시행으로 인해 검사에서 재고자 하는 특성과 더불어 다른 요인들의 영향이 각 학년의 학생들에게 영향을 미칠 수 있으며, 이로 인해 IRT의 일차원성 가정이 위배될 가능성이 있다. 따라서 척도검사설계방법을 통해 수직척도를 개발하기 위해서는 검사 시행, 데이터 클리닝, 문항모수 추정 등에 있어서 주의를 요해야 하며, 특히, 일차원성 가정의 검토는 필수적으로 수행되어야 할 것이다. 이 연구에서 산출된 RMSE, BIAS, SEE의 수치들은 평균이 0이며 표준편차가 1인 θ점수 체제에서 산출되었으며, 학생들의 개별적인 점수 차이보다는 전체적인 분포의 모수를 비교하였기 때문에 RMSE, BIAS, SEE의 수치들이 작은 값으로 산출되었다. 하지만 일반적으로 IRT 수직척도를 개발하여 실제적으로 적용을 할 때에는 θ점수를 그대로 보고하

기보다는 이를 일반인들이 이해하기 쉬운 척도로 변환한 척도점수를 보고한다. 따라서, 이러한 수치들이 수직척도가 실제 적용될 때 어떤 의미를 가지는 지는 보고되는 척도점수에 따라 달라질 수 있다. 척도점수로 변환하는 방법은 여러 가지가 있지만 척도점수로 변환하는 과정에서 또 다른 오차가 개입할 수 있으며 이는 개발된 수직척도에서 보고되는 척도점수에 영향을 미칠 수 있다. 이러한 요인으로 인해 실제 수직척도를 개발하여 보고하는 척도점수의 변환방법에 대한 추가적인 연구가 시행되어야 하며 지역독립성 가정의 위배와 관계에 대한 추가적인 연구가 필요하다.

참 고 문 헌

김성훈(2000). 능력 변화과정 추정을 위한 시험자료의 동등화 방법 연구. **교육평가연구, 13**(1), 101-125.

민경석(2010). 수직 척도점수를 이용한 학년 간 비교의 타당성 - 초등학생용 ICT 리터러시 검사. **교육과정평가연구, 13**(1), 25-42.

박인용, 이규민, 강상진(2012). IRT 수직척도의 자료수집방법과 추정방법에 따른 피험자 능력점수의 성장패턴과 척도변동성. **교육평가연구, 25**(2), 263-286.

부재율(2005). 가교문항 수직적 동등화를 이용한 국어, 영어, 수학 학업성취도의 학년 간 비교 연구. **교육평가연구, 18**(1), 127-152.

이규민, 강상진, 노명완, 유제명, 류희찬(2006). 국가수준 종단적 교육조사 연구를 위한 성취도 검사 및 척도개발. 한국교육과정평가원 연구보고 RR 2006-1.

이규민, 임현정, 박인용, 김연정(2010). 「한국교육종단연구 2005」의 수직척도 타당성 검토. **교육평가연구, 23**(3), 617-640.

Andrews, K. M. (1995). *The effects of scaling design and scaling method on the primary score scale associated with a multi-Classification Proportion Level achievement test*. Unpublished Doctoral Dissertation, The University of Iowa, Iowa City.

Bishop, N. S., & Omar, M. H. (2002). *Comparing vertical scales derived from dichotomous and polytomous IRT models for a test composed of testlets*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3-14.

Camilli, G. (1988). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics, 13*(3), 227-241.

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement, 36*(1), 73-78.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379-388.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. NY: Guilford Press.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. New York: Springer-Verlag.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory, principles and applications*. Boston: Kluwer.

Hendrickson, A. B., Kolen, M. J., & Tong, Y. (2004). *Comparison of IRT vertical scaling from scaling test and common item designs*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Hendrickson, A. B., Wei, H., Kolen, M. J., & Tong, Y. (2005). *Dichotomous and polytomous scoring for IRT vertical scaling from scaling-test and common-item designs*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice, 3*(4), 8-14.

Kim, J. (2007). *A comparison of calibration methods and proficiency estimators for creating IRT vertical Scales*. Unpublished Doctoral Dissertation, The University of Iowa, Iowa City.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*(1), 1-11.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. (2nd ed.). New York: Springer-Verlag.

Lee, G. (2000). A comparison of methods of estimating conditional standard errors of measurement for testlet based test scores using simulation techniques. *Journal of Educational Measurement, 37*(2), 91-112.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*(3), 237-255.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice, 19*(4), 5-9.

Lee G., & Park, I.-Y. (2008). *A comparison of the approaches of classical test theory, generalizability theory, and item response theory in estimating the reliability of test scores composed of testlets*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Lee, G., Park, I.-Y., Jeon, M.-J. (2009). Testlet response model for IRT true score equating. *Journal of Educational Evaluation, 22*(3), 871-887

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss(Ed), *New horizons in testing,* 147-176. New York: Academic.

Meng, H. (2007). *A comparison study of IRT calibration methods for Mixed-format tests in vertical scaling*. Unpublished Doctoral Dissertation, The University of Iowa, Iowa City.

Mislevey, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software, Inc.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*(2), 99-117.

Ngudgratoke, S. (2006). *Vertical linking of test composed of testlets: A comparison between dichotomous model, polytomous model, and testlet response model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Reckase, M. D., & McKinley, R. L. (1983). *The definition of difficulty and discrimination for multidimensional item response theory models*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495-529.

Tong, Y. (2005). *Comparison of methodologies and results in vertical scaling for educational achievement tests*. Unpublished Doctoral Dissertation, The University of Iowa, Iowa City.

Tong, Y., & Kolen, M. J. (2006). *Vertical scaling and scale shrinkage*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admission test as an example. *Applied Measurement in Education, 8*(2), 157-186.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3-PL useful in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice,* 245-270. Boston, MA: Kluwer-Nijhoff.

Wang, W.-C., & Wilson, M. R. (2000). Using a new statistical model for testlets to score

TOEFL, *Journal of Educational Measurement, 37*(3), 203-220.

Wang, W.-C., & Wilson, M. R. (2005). Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126.

Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*(2), 93-107.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 299-325.

Zimowski, M. F., Muraki, E., Mislevy, R. J., and Bock, R. D. (2003). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software.

• *논문접수 : 2013-05-01/ 수정본접수 : 2013-06-10/ 게재승인 : 2013-06-19*

ABSTRACT

The effects of violation of the local independence assumption on scale variability in IRT vertical scale

In-Yong Park
(Associate Research Fellow, Korea Institute for Curriculum and Evaluation)

When IRT is used to construct a vertical scale, there are many effects of many factor which are model, assumption, data collection design on vertical scale. The purpose of this study is to examine the effect of violation of local independence assumption on scale variability in IRT vertical scales for tests composed of testlet based on different data collection designs. Data collection designs used in this study were scaling test design and common item design. The data were generated in two dimensional compensatory model incorporated testlet effect. In result, when the local independence assumption was satisfied, scale variability yielded different patterns between scaling test design and common item design. Scale variability under both of data collection design showed scale shrinkage when the local independence assumption was not met. And the magnitude of shrinkage in common item design was more than those in scaling test design. Therefore, in practical, when IRT is used to construct a vertical scale with sever violation of local independence assumption, it must be controlled testlet effect using testlet-based scoring or testlet model incoporated testlet effects.

Key Words : local independence assumption, item response theory, vertical scale, scale variability, common item design, scale test design