교육과정평가연구
The Journal of Curriculum and Evaluation
2008, Vol. 11, No. 2, pp. 71～87

컴퓨터를 활용한 변형 선다형 검사의 차원성[1)]

민 경 석(세종대학교 조교수)
박 주 용(세종대학교 부교수)

≪ 요 약 ≫

학교 검사와 표준화 검사에서 주로 활용되는 선다형 문항은 다양한 장점(내용 타당도, 검사 점수의 신뢰도 등)에도 불구하고 고등사고능력을 측정하기 어렵다는 비판을 받아왔다. 반면에 구성형 문항은 내용의 포괄성, 채점의 신뢰도라는 측면에서 비효율적이라는 문제점을 보인다. 교육평가 영역에서 컴퓨터 기술의 활용은 전통적 선다형 문항과 구성형 문항의 단점을 보완하고 장점을 결합한 새로운 문항 형식을 모색하여 왔다. 대표적으로 변형 선다형 검사는 피험자가 단답형과 같이 정답을 숙고하고 선다형 선택지를 이용하여 최종 반응을 하게 하는 검사 방식이다. 컴퓨터를 활용한 새로운 검사 방식의 측정이론적 구인의 특성을 밝히고자 초등학교 6학년 학생을 대상으로 변형선다형 검사와 전통적 선다형 검사를 실시하였으며, 자료 분석을 위하여 다차원 문항반응 모형과 군집분석을 적용하였다. 연구결과로써 변형 선다형 검사와 전통적 선다형 검사의 측정 구인을 비교하였으며, 후속연구를 위한 제안을 논의하였다.

주제어 : 구성형 문항, 변형선다형 검사, 선다형 문항 , 다차원 문항반응모형, 군집 분석

Ⅰ. 컴퓨터를 활용한 새로운 문항 형식

학생들의 인지적 능력을 평가하는 교사제작 검사(teacher-made assessment)와 표준화 검사(standardized test)는 일반적으로 선다형 문항(selected response items)으로 구성된다. 선다형 문항은 채점의 간편성과 공정성이라는 장점을 가질 뿐만 아니라 문항당 풀이 시간이 상대적으로 짧아 제한된 시간 동안 많은 문항을 풀게하여 내용 타당도(content validity)를 갖출 수 있다(Wainer & Thissen, 1993). 그러나 이러한 장점에도 불구하고 선다형 문항은 학생들이 자신

1) 이 논문은 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2007-B00130).

의 생각을 표현하기 보다는 답지를 선택한다는 수동적 측면에서 고차원적 사고를 측정하기 어렵다는 비판을 받아왔다(Sireci & Zenisky, 2006). 특히 검사를 통하여 '안다'의 차원을 넘어 '행한다'를 측정하는 수행평가에서 선다형 문항보다 직접 학생이 자신의 생각을 정리하고 표현하는 구성형 문항(constructed response items)의 중요성이 지속적으로 강조되어 왔다. 피험자가 문제해결과정을 진술하고 이에 대한 결론을 주장하는 구성형 문항은 실제 학생의 사고과정, 논리적 표현을 직접 측정한다는 점에서 중요한 의미를 갖는다. 그러나 구성형 문항으로 구성된 검사는 측정이론적 측면에서 낮은 신뢰도와 적은 문항수로 인한 내용적 비포괄성 등의 점에서 비판을 받아왔으며, 실용적인 측면에서도 검사시간이 길어지며, 점수 산출을 위한 인적/시간적 비용이 크다는 점에서 현실적 적용에 어려움을 나타낸다(Wainer & Thissen, 1993).

근래 컴퓨터 기술의 발달과 대중적 활용은 다양한 형태로 교육현장에 영향을 미쳐왔다. 특히 교육평가영역에서 컴퓨터 기술은 전통적인 선다형 문항의 제한성을 극복할 수 있는 다양한 형태의 문항 형식을 활용할 수 있는 가능성을 현실화시켰다. 컴퓨터를 활용한 새로운 형식의 문항(innovative items)은 측정 내용과 채점절차라는 두 가지 측면에서 그 유용성을 평가할 수 있다. 먼저 측정 내용에 있어 컴퓨터 활용 문항은 다양한 반응(복수선택, 관계설정, 정보 순서화, 분류, 문장수정, 문장완성, 그림모형 등)을 유도하여 전통적인 선다형 문항이 측정하기 어려운 지식, 기술, 능력을 측정한다(Sireci & Zenisky, 2006; Zenisky & Sireci, 2002). 둘째, 컴퓨터를 활용한 문항은 피험자의 반응이 컴퓨터에 저장되고 사전에 설정된 논리적 연산 절차에 따라 정답 유무, 할당 점수 등이 즉각적으로 산출되어 사람이 채점과정에 개입하여 나타날 수 있는 오류 및 비용을 최소화한다.

교육평가 영역에서 컴퓨터 기술 도입의 성과인 새로운 문항 형식 개발은 전통적 선다형 문항과 구성형 문항의 단점을 보완하고 장점을 결합하는 가능성을 모색하여 왔다. 대표적으로 Park(2005, 2007), Park과 Choi(2008) 등이 개발/적용한 변형 선다형 문항(modified multiple-choice items)이 구체적 사례라고 할 수 있다. 변형 선다형 문항은 피험자가 단답형과 같이 가능한 정답을 생각하고 실제 응답은 선택지에서 정답을 선택한다는 점에서 전통적 선다형 문항과 구성형 문항의 특성을 결합한 문항 형태라고 할 수 있다. 특히 피험자가 단답형 문항에서와 같이 정답을 숙고한다는 점에서 문항 추측도를 제거하는 장점을 가지며, 최종적으로 선택지에서 정답을 선택한다는 측면에서 채점의 간편성을 유지한다.

전통적 선다형 검사의 선택지 형식(순서, 개수 등)에 따른 측정구인의 변화에 대한 다양한 논의를 고려할 때(Rodrigues, 2003, 2005; Wainer & Thissen, 1993), 컴퓨터를 활용한 새로운 문항형식을 활용한 기존 연구에 대한 주요한 비판점은 대부분 문항 형식 개발에 치중한 나머지, 새로 개발된 문항 형식이 측정하는 바가 무엇이며 기존 문항 형식과 어떤 다른 점이 있는가를 명확히 규명하지 못한다는 것이다(Huff & Sireci, 2001). 즉, 전통적 문항형식과 비

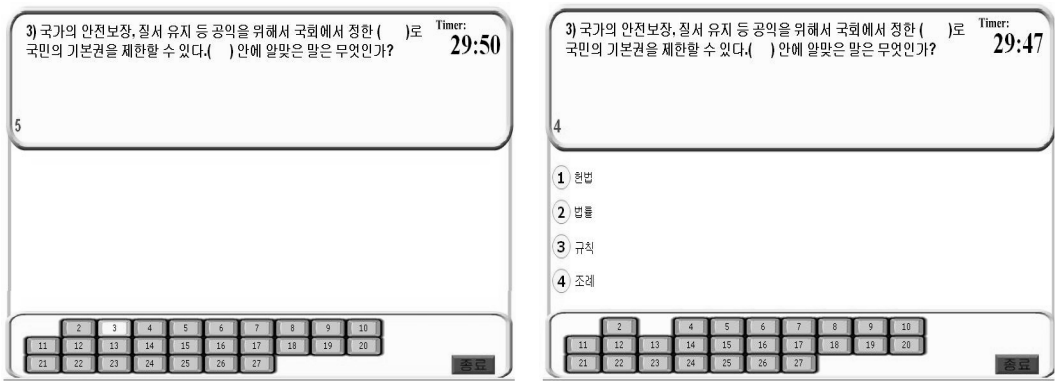
교하여 새로운 문항 형식이 측정하는 구인(construct)이 어떤 차별성과 동질성을 갖는지에 대한 규명이 새로운 문항 형식의 활용과 개선을 위한 선행 조건이라 할 수 있다. 특히 컴퓨터를 이용한 검사의 경우 전통적인 지필검사와는 매우 다른 환경에서 검사가 실시되며, 추가적으로 문항형식 또한 새로운 형태라는 점에서 전통적 선다형 검사의 측정 구인과의 비교는 중요한 의미를 갖는다.

이 논문에서는 구성형 문항과 선다형 문항 형식을 결합한 컴퓨터 활용 변형 선다형 검사에 대하여 논의하고, 다차원 문항반응모형(multidimensional IRT models)과 군집분석(cluster analysis)의 문항군집 구조 분석을 통하여 변형 선다형 문항의 측정 구인 특성을 전통적 선다형 문항과 비교했다.

Ⅱ. 변형 선다형 검사

컴퓨터를 이용한 변형선다형 검사, CMMT(computerized modified multiple-choice test)는 구성형 문항 형식인 단답형을 풀도록 한 다음 선택지를 이용하여 최종 정답을 선택하게 하는 검사 방식이다(Park, 2005). 이 방식은 선다형 문항에서 사고를 수동적으로 만들거나 답지를 활용한 추측의 문제점을 해결하고자 한다.

구체적으로 변형선다형 문항의 시행절차를 몇 개의 단계로 나누어 설명하면 다음과 같다(〔그림 1〕 참조). 먼저 피험자들이 화면 하단의 문항 번호에 마우스를 갖다 대면 화면에 답지가 없이 문제만을 제시하되, 정해진 시간 동안에는 문제를 얼마든지 볼 수 있다. 둘째, 특정한 문항(예를 들면 3번 문항)에 대해 피험자가 답을 스스로 찾았다고 생각하면(단답형으로 정답 추정), 마우스를 클릭하여 선택지를 제시하도록 요구한다. 선택지는 미리 정해진 짧은 시간만큼(선택지의 길이나 이해도에 따라 달라짐, 3번 문항의 경우 5초) 제시되며, 피험자는 그 시간 내에서만 반응할 수 있다(선다형으로 정답 선택). 셋째, 정해진 시간이 지나거나 피험자가 정답을 결정하면 선택지와 함께 문항 자체가 화면에서 사라져 더 이상 반응을 할 수 없게 된다. 이러한 변형선다형 문항의 핵심은 피험자가 기본적으로 단답식처럼 문제를 풀고, 마지막 순간에 선택지를 이용하여 자신이 생각한 정답을 골라 반응하게 하는 것이다.



(a) 문항 제시 화면 (b) 선택지 제시 화면

〔그림 1〕 변형 선다형 문항 예시

(a) 피험자가 1번 문항을 이미 풀었기 때문에 화면 하단의 1번 박스는 삭제되었음. 마우스를 하단 3번 박스에 올리면 3번 문항이 선택지 없이 제시됨. 화면 좌측 중상단의 숫자(5)는 선택지가 제시될 시간(5초)을 나타냄. (b) 피험자가 마우스를 하단 3번 박스를 클릭하면 선택지가 제시되며, 시간 경과(4)가 화면에 표시됨. 주어진 5초 동안 피험자는 선택지에서 정답을 선택/변경할 수 있음. 5초가 지나면 화면에서 문제와 답지가 사라지며 더 이상 반응할 수 없음.

변형 선다형 검사 방식을 통해 기대되는 효과는 우선 단답형에서처럼 인출의 난이도를 증가시킬 수 있다는 것이다(예, Bjork, 1994; Carrier & Pashler, 1992). Carrier와 Pashler는 대학생들에게 40쌍의 무의미 철자와 숫자(예, KIR-35)를 학습하도록 하였다. 한 집단에는 쌍으로 된 자극을 10초 동안 제시하였고, 다른 집단에는 처음 5초 동안 각 쌍의 앞부분만(예를 들면, KIR) 제시하고 나머지 5초 동안은 쌍으로 제시하였다(즉, KIR-35). 그 결과 무의미 철자와 숫자 쌍이 모두 함께 10초 동안 제시된 집단보다, 인출이 촉진되도록 5초 동안은 무의미 철자만이 나머지 5초 동안은 무의미 철자와 숫자 쌍이 함께 제시되었던 집단의 수행이 더 높음을 발견하였다. 이런 맥락에서 Park(2005)은 문제만 먼저 제시되는 변형 선다형 방식이 전통적 선다형 방식에서보다 인출이 더 촉진되는가를 검증하는 실험을 수행하였다. Park은 초등학교 6학년 학생을 두 집단으로 나누어 한 집단은 변형 선다형 방식으로 다른 집단은 전통적 선다형 방식으로 시험을 보게 하였다. 며칠 후 지필식으로 본 최종 시험에서 변형선다형 방식으로 중간시험을 본 집단이 전통적 선다형 방식으로 본 집단보다 단답식으로 본 최종시험 점수가 더 높음을 발견하였다. 후속 연구(Park & Choi, 2008)에서 컴퓨터로 보는 시험을 집에서 보게 했을 때에도 CMMT방식으로 볼 때가 전통적 선다형 방식보다 최종 시험에서 더 높은 점수를 받는 결과를 얻었다. 이상의 결과는, CMMT가 전통적 선다형보다 인출의 난이도를 어렵게 하여 나중에 다시 기억해 낼 때 기억을 더 고양시켰음을 의미한다.

CMMT 방식은 선택지를 짧게 제시하기 때문에 선택지만 가지고 추측하여 정답을 결정하는 것을 어느 정도 배제할 수 있다. 또한 반응 자체는 선다형처럼 선택지를 고르는 것이기에 채점이 쉽고 객관성을 유지할 수 있다는 점, 그리고 선택지를 만드는 일이 상대적으로 쉬워져 출제가 용이해진다는 점도 장점이라 할 수 있다.

Ⅲ. 검사의 차원성 : 다차원 문항반응모형과 위계적 군집분석

측정의 결과로써 다양한 방식으로 산출된 '검사 점수가 의미하는 바가 무엇인가?'라는 질문은 측정이론의 중요한 관심 사항이라 할 수 있다. 일반적으로 교육/심리 검사는 단일한 차원보다는 여러 차원의 능력 혹은 특성과 관계된 것으로 인식되어왔으며(Ackerman, 1994), 이를 다차원 문항반응 모형으로 표현하면 식 (1)과 같다(민경석, 2004; Reckase, 1985, 1991).

$$P(u_{ij}=1|\mathbf{a_i}, \mathrm{d_i}, \boldsymbol{\theta_j}) = \frac{\exp\left(\sum_{k=1}^{m} \mathrm{a_{ik}}\theta_{\mathrm{jk}} + \mathrm{d_i}\right)}{1+\exp\left(\sum_{k=1}^{m} \mathrm{a_{ik}}\theta_{\mathrm{jk}} + \mathrm{d_i}\right)} \quad (1)$$

여기서 u_{ij}는 j번째 피험자의 i번째 문항에 대한 반응(1 정답; 0 오답), $\boldsymbol{\theta_j} = (\theta_{j1}, \theta_{j2}, \cdots, \theta_{jm})'$는 j번째 피험자의 능력모수 벡터, $\mathbf{a_i} = (\mathrm{a_{i1}}, \mathrm{a_{i2}}, \cdots, \mathrm{a_{im}})'$는 i번째 문항의 변별도 벡터, d_i는 i번째 문항의 난이도와 관계있는 위치모수, m은 능력차원의 개수를 가리킨다.

다차원성에 근거하여 벡터(vector)형태를 갖는 다차원 문항반응모형의 모수는 일차원 문항반응모형의 모수와 직접 비교될 수는 없으나 Euclid 기하 공식을 이용한 문항 모수(다차원 문항 변별도 MDISC, 다차원 문항 난이도 MDIFF)를 통하여 두 모형의 모수간 상호비교가 가능하다(Reckase, 1985). 특히 다차원 문항변별도는 잠재능력 공간에서 정답 확률의 변화 정도가 가장 큰 지점에서의 기울기에 비례하며 그 방향은 식 (2)로 나타낼 수 있다.

$$\cos(\alpha_{ik}) = \frac{\mathrm{a_{ik}}}{MDISC_i} \quad (2)$$

식 (2)에서 α_{ik}는 k번째 차원과 문항벡터(item vector)가 이루는 각의 크기를 나타내며, 여기서 문항 i와 i′가 측정하는 차원의 유사성은 식 (3)의 각거리(angular distances, $\alpha_{ii'}$)로 표현된다(Miller, Hirsch, 1992).

$$\cos(\alpha_{ii'}) = \frac{\mathbf{a}'_{\mathrm{ii'}} \bullet \mathbf{a}_{\mathrm{ii'}}}{|\mathbf{a}_{\mathrm{ii'}}|\ |\mathbf{a}_{\mathrm{ii'}}|} \quad (3)$$

식 (3)에서 각거리 $\alpha_{ii'}$는 문항 벡터 i와 i′사이의 각도(degree)를 나타내는 것으로 0에 가까울수록 두 문항이 같은 차원을, 90°에 가까울수록 서로 독립적 차원을 측정하는 것을 의미한다. 곧 각거리 $\alpha_{ii'}$는 서로 다른 문항군집을 구별하는 차별성(dissimilarity) 값으로 군집분석의 입력자료로 활용될 수 있다[2].

유사한 특성을 보이는 사례나 변수를 통계적 집단으로 묶어 주는 군집분석은 검사의 차원성이 많아질 때 측정하는 바를 명시화하는 방법으로 다양하게 사용되어왔다(Miller & Hirsch, 1992; Roussos, Stout, & Marden, 1998). 이 연구에서는 다양한 군집분석의 방법 중 선행연구의 제안에 따라 군집 결정 절차로써 완전연계방법[3](complete linkage method, Aldenderfer & Blashfield, 1984; Milligan & Cooper, 1985)과 대상의 차별성을 나타내는 입력자료로써 각거리를 활용하였다.

Ⅳ. 연구방법

1. 검사 대상 및 내용

전통적 선다형 문항과 변형 선다형 문항을 활용한 검사의 결과를 비교하기 위하여 서울에 위치한 두 개 초등학교(J와 P 초등학교)에서 컴퓨터 과목을 방과후 활동으로 참여한 6학년 재학생 374명에 대하여 검사를 실시하였다. J학교의 7개 학급 198명과 P학교 6개 학급이 대상이 되었으며, 전체 374명 중 여학생 179명, 남학생이 195명이었다. 총 13개 학급은 학교 별로 두 집단으로 나뉘어 각각 전통적 선다형 검사 집단(175명)과 변형 선다형 검사 집단(199명)에 임의할당(random assignment)되었다.

검사 문항은 초등학교 6학년 사회과목(사회 6-2, 사회과 탐구 6-2)으로 실험 직전 2주 동안 학습한 내용으로 구성되었으며, 재직 경력 10년 이상인 2명의 초등학교 교사가 문항을 개발하였다. 전체 검사는 27문항(20개 고유문항, 7개 공통문항)[4]으로 구성되었으며, 검사시간은 30분으로 제한하였다.

---

2) 유사한 대상/변수를 묶어주는 군집분석은 일반적으로 자료의 특성에 따라 유사성(similarity) 혹은 차별성(dissimilarity)을 나타내는 통계치를 분석자료로 이용한다(Aldenderfer & Blashfield, 1984).

3) 이외에 단순연계법(single linkage), 평균연계법(average linkage) 등의 방법이 있으나 완전연계법은 군집내 모든 대상의 특성을 군집 결정에 활용한다는 점에서 군집결정이 상대적으로 안정적이라 할 수 있다.

4) 고유문항과 공통문항에 대한 일반적 정의와는 달리 여기서 고유문항은 문항 내용이 정확히 동일하지만 선다형 혹은 변형 선다형이라는 문항 형식에서 차이가 있음을 나타내고, 공통문항은 문항 내용뿐만 아니라 문항 형식도 선다형으로 일관됨을 의미한다.

2. 검사 시행

고유문항과 공통문항으로 구성된 검사는 두 가지 방식으로 실시되었다. 먼저 선다형 집단에서는 고유문항과 공통문항이 모두 전통적인 선다형 방식으로 실시되었다. 반면에 변형 선다형 집단에서는 20개 고유문항은 변형 선다형 방식으로 7개 공통문항은 전통적 선다형 방식으로 실시되었다.

선다형 집단과 변형 선다형 집단 모두에게 컴퓨터를 이용하여 검사를 실시하였다. 다만 선다형 집단의 경우 27개 문항 모두에서 질문과 선택지가 함께 제시되었으며, 변형선다형 집단의 고유문항인 20문항은 문제가 먼저 제시되고 이후 피험자가 요구하는 경우(마우스 클릭) 선택지가 짧은 시간동안 제시된다.

3. 자료 분석

학급을 단위로 임의할당을 실시하였음에도 불구하고 선다형 집단과 변형 선다형 집단의 사전 동등성을 확인하기 위하여 동일한 문항 형식(선다형)으로 시행된 공통문항(7문항)에 대한 정답률을 비교했다. 일반적으로 선택지가 짧은 시간 동안 제시되는 변형 선다형 문항의 정답률은 낮아지는 경향을 보이기 때문에 동일한 내용과 동일한 문항 형식을 갖는 공통 문항이 두 집단의 동등성을 비교하는 가장 적절한 기준으로 활용될 수 있다.

정확히 동일한 내용인 선다형 문항과 변형 선다형 문항이 측정하는 바가 동일한지를 검증하기 위하여 6차원 2모수 문항반응 모형과 위계적 군집분석 방법을 이용하였다. 먼저 다차원 문항반응 모형에서 6차원을 설정한 것은 실제 검사가 측정하는 바를 보다 포괄적으로 확인하기 위하여 상대적으로 많은 차원을 설정하였으며(Hirsh & Miller, 1991; Reckase & Hirsh, 1991), 적은 사례수(피험자 수)를 고려하여 추측요인을 제외한 2모수 모형을 활용하였다(식 (1) 참조). 검사의 차원이 고차원으로 갈수록 수리적, 개념적으로 검사가 측정하는 바에 대한 명확한 이해가 어렵기 때문에 다차원 문항반응모형으로 추정된 문항 특성(변별도, 각거리)을 입력자료로 활용하여 위계적 군집분석(Miller & Hirsch, 1992)을 실시하였다.

Ⅴ. 연구결과

1. 집단 동등성 및 검사 신뢰도

검사점수는 피험자와 문항의 상호작용이라 할 때 선다형 문항 집단과 변형 선다형 문항 집단에서 검사의 차원성을 비교하기 전에 피험자 집단이 동질적임을 판단할 필요가 있다. 피험자의 동질성 비교를 위하여 임의할당을 통하여 학급을 두 집단에 배치하였다. 그러나 우연적 요인에 따라 집단간 차이 가능성을 확인하기 위하여 선다형 문항형식의 7개 공통문항을 통하여 집단간 차별성을 비교하였다.

〈표 1〉 공통 문항의 정답률과 총점 비교

문항 번호	변형선다형	선다형	차이	t
공통문항 1	0.82	0.79	0.03	.60
공통문항 2	0.59	0.61	-0.03	-.52
공통문항 3	0.69	0.73	-0.04	-.86
공통문항 4	0.73	0.72	0.01	.16
공통문항 5	0.21	0.19	0.02	.44
공통문항 6	0.54	0.58	-0.03	-.63
공통문항 7	0.25	0.28	-0.03	-.74
공통문항 총점	3.83	3.92	-0.09	-.58
사례수	188	165	-	-

〈표 1〉에 제시된 바와 같이 변형 선다형 문항 집단과 선다형 문항 집단 사이에 공통문항에 대한 정답률 차이가 통계적으로 유의미하지 않은 것으로 나타났다. 선다형 집단이 전체적으로 변형 선다형 집단 보다 공통문항 총점에서 약간 높은 점수를 나타내고 있으나, 이 또한 통계적으로 유의미하지 않았다. 즉, 선다형으로 구성된 공통문항에 있어 두 집단 사이에 능력 차이는 통계적으로 유의하지 않았다.

집단간 능력 동등성을 확인한 후 두 가지 검사 점수 신뢰도(Cronabch's alpha)를 산출하였으며, 부가적으로 각 집단에서 공통문항 점수와 고유문항 점수의 상관을 비교하여 〈표 2〉에 제시하였다.

〈표 2〉 신뢰도와 상관계수

집단	검사 전체(27문항) 신뢰도	고유문항(20문항) 신뢰도	상관계수 (공통문항, 고유문항)
변형선다형	0.72	0.69	0.43**
선다형	0.69	0.66	0.33**

* p<.05, ** p<.01

일반적으로 학교 검사의 신뢰도가 0.7 내외인 점을 고려할 때, 이 연구에서 실시된 두 검사는 모두 적절한 수준이라고 할 수 있다. 그러나 변형 선다형 집단과 선다형 집단 모두의 상대적으로 낮은 상관계수는 공통문항과 고유문항이 엄밀할 의미의 동형관계 수준에 있지 못함을 나타낸다.[5)]

〈표 2〉에서 주목할 점은 변형선다형 집단의 검사 점수 신뢰도가 전체 검사와 고유문항 모두에서 약간 높게 나타난 것으로, 이는 변형선다형 검사의 신뢰도가 구성형 검사의 문제점인 낮은 신뢰도와 달리 선다형 검사의 신뢰도 수준을 확보하고 있음을 의미한다.

2. 다차원 문항반응 모형

동일한 내용에 대하여 문항 형식 차이에 따른 검사 구인의 차원성 비교를 위하여 식 (1)의 다차원 문항반응 모형을 변형 선다형 검사와 선다형 검사에 각각 적용하였으며, 추정된 문항 변별도가 〈표 3〉에 제시되었다.

문항변별도 추정을 위하여 각 검사에 6개 차원이 설정되었으며, 문항 추측도를 제외한 2모수 모형이 적용되었다. 여기서 6개 차원이 설정된 것은 검사 점수의 다양한 의미를 충분히 포괄하기 위하여 상대적으로 많은 차원성이 고려된 것이며, 문항 추측모수는 검사 점수의 차원성과는 관련이 없는 것으로 사례수를 고려하여 상대적으로 간편한 모형을 적용하였다[6)].

5) 이 연구의 목적이 공통문항을 통한 검사 동등화에 있지는 않지만, 공통문항에 대한 세부적인 분석을 통하여 변형선다형 집단과 선다형 집단 모두의 공통문항 점수 분포에서 상대적으로 작은 분산이 나타났으며(중간 난이도와 낮은 변별도), 이러한 작은 분산이 공통문항 점수와 고유문항 점수 간의 낮은 상관계수로 이어졌다.

6) 실제 추측모수를 고려한 3모수 문항반응모형을 적용한 결과 문항 변별도는 〈표 3〉과 일정정도 다른 값으로 추정되지만, 군집분석 결과는 동일하게 나타남.

〈표 3〉 다차원 문항반응모형의 문항 변별도

문항	변형선다형 검사						선다형 검사					
	a1	a2	a3	a4	a5	a6	a1	a2	a3	a4	a5	a6
1	0.94	0.00	0.00	0.00	0.00	0.00	0.77	0.00	0.00	0.00	0.00	0.00
2	0.33	1.11	0.00	0.00	0.00	0.00	0.12	1.06	0.00	0.00	0.00	0.00
3	0.65	0.07	0.66	0.00	0.00	0.00	0.53	-0.14	0.8	0.00	0.00	0.00
4	0.05	0.57	0.05	0.32	0.00	0.00	-0.3	1.05	1.01	0.27	0.00	0.00
5	-0.03	0.1	0.48	0.52	0.47	0.00	0.69	0.07	0.1	0.33	0.42	0.00
6	-0.07	0.2	0.11	0.54	-0.12	0.04	0.11	0.52	-0.03	0.5	0.08	0.06
7	0.15	-0.02	0.15	0.29	0.92	-0.76	0.53	0.01	0.45	0.35	0.46	0.27
8	0.33	-0.02	-0.24	0.96	0.18	0.37	-0.19	0.33	0	0.32	-0.06	0.39
9	-0.36	-0.21	-0.13	0.18	0.6	-0.14	-1.82	-5.27	5.07	1.34	2.25	7.95
10	-0.2	0	-0.6	0.49	0.17	0.2	-0.3	0.18	-0.28	0.27	-0.07	0.8
11	-0.56	-0.17	0.03	-0.12	0.63	0.32	0.27	-0.16	-0.05	-0.22	0.11	0.29
12	-0.05	0.03	-0.19	-0.04	0.27	0.44	0.01	-0.21	-0.23	-0.11	-0.86	0.43
13	0.12	0.04	-0.25	-0.59	1.03	-0.23	0.31	0.12	-0.21	-0.33	0.19	0.5
14	0.23	0.22	-0.06	-0.32	0.24	0.47	0.32	0.1	-0.05	-0.12	-0.62	0.2
15	0.28	0.34	0.01	-0.08	0.46	-0.22	0.52	0.51	0.09	-0.53	0.04	0.32
16	0.27	0.32	0.15	0.05	-0.01	0.23	0.24	0	0.52	0.48	-0.84	-0.17
17	0.05	0.45	0.14	0.14	0.19	-0.31	0.05	0.31	0.97	0.13	-0.06	0.29
18	0.18	0.17	0.54	0.16	0.27	0.58	0.44	-0.37	0.41	0.92	0.14	-0.12
19	-0.44	0.36	0.41	0.27	0.29	0.09	-0.18	0.2	1.03	-0.08	0.32	0.02
20	-0.01	0.1	0.05	0.5	0.07	0.44	-0.04	-0.1	0.05	0.31	-0.1	0.07

다차원 문항 변별도의 추정을 위해서 NOHARM 프로그램(Fraser, 1986)이 이용되었다. NOHARM을 통해 추정되는 차원의 방향성(dimensional directions, Li & Lissitz, 2000)은 임의적인 것으로 〈표 3〉에 제시된 바와 같이 1번 문항은 첫 번째 차원(a1)에만 반응하고, 2번 문항은 첫 번째(a1)와 두 번째 차원(a2)에 적용되는 순으로 문항 변별도가 추정된다. 이와 같이 차원의 방향성이 임의적으로 설정됨에도 불구하고 문항벡터의 상대적 유형과 이에 따른 문항 군집은 문항 제시 순서에 영향을 받지 않는다(Miller & Hirsch, 1991).

검사의 차원성을 분석하기 위한 문항 변별도의 해석 방법은 요인 분석의 부하량(factor loadings)과 유사한 것으로 변별도의 절대값이 높을수록 해당 차원과 관련성이 높다는 것을 나타낸다. 예를 들어 문항 7은 변형 선다형 검사에서 다섯 번째 차원에 가장 높은 변별력을 보이는(a5=0.92) 반면, 선다형 검사에서는 첫 번째 차원에서 가장 높은 변별력을 보인다

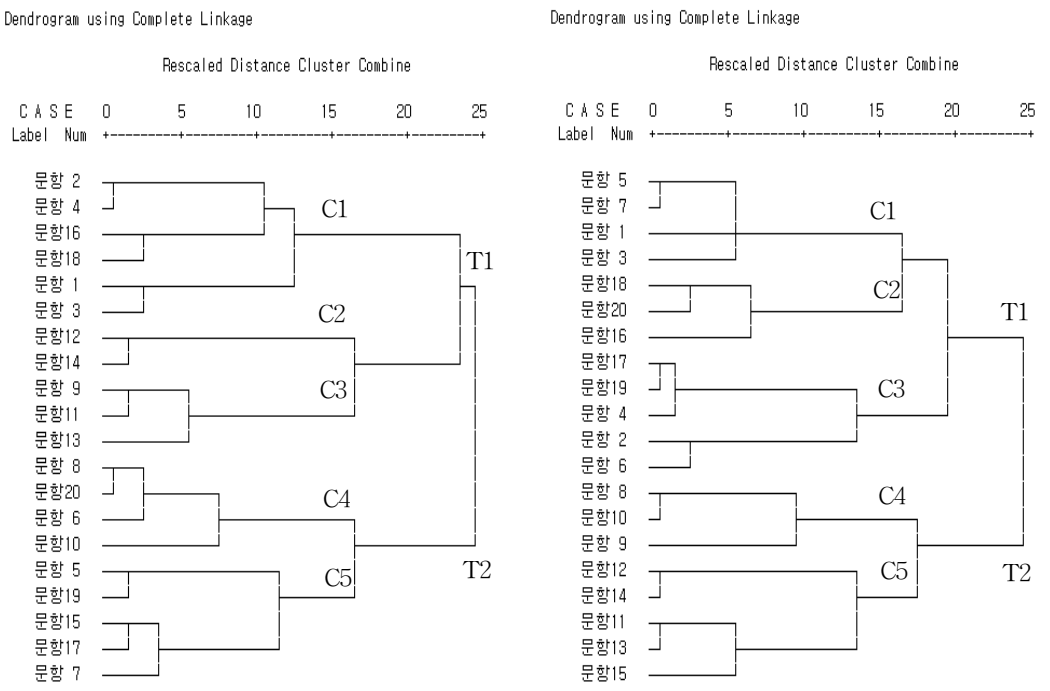
(a1=0.53). 이와 같이 정확히 동일한 내용으로 구성된 문항임에도 문항 형식에 따라 일정한 차별성이 나타나고 있음을 〈표 3〉에서 확인할 수 있다.

3. 위계적 군집 분석

다차원 문항반응모형으로 추정된 변별도는 검사의 차원성을 확인하는 데 중요한 자료로 활용된다. 그러나 검사의 차원수가 많아질수록 문항 변별도를 이용한 차원성에 대한 분석/해석이 어려워진다. 실제 다차원 문항반응이론에 관한 선행연구에서 다양한 그림, 그래프를 이용하여 보다 의미있는 차원성 해석을 시도하였으나 대부분의 연구는 2차원, 3차원 수준으로 제한된다. 이러한 점을 해결하기 위하여 군집분석이 제안되었으며(Miller & Hirsch, 1991), 군집분석은 대상의 특성이 나타내는 유사성과 차별성에 근거하여 집단을 구분한다는 점에서 검사 차원성을 해석하는 데 도움을 준다. 특히 군집분석은 문항 내용보다 정답률에 따른 측정 구인 설정을 방지한다는 점에서 긍정적 특성을 갖는다.

〈표 3〉의 변별도 추정치와 식 (2), (3)을 활용한 위계적 군집분석 결과로써 위계그림(dendrogram)이 〔그림 2〕에 제시되었다. 그림에 제시된 바와 같이 문항 내용이 동일한 20개 고유문항은 차원성(군집)이라는 측면에서 변형선다형과 선다형 검사는 일정한 차별성을 보인다.[7] 포괄적인 수준에서 두 검사는 두개의 문항 군집(T1, T2)으로 구성된다. 먼저 변형 선다형 검사는 지식/이해 수준의 문항군집과 적용 수준의 문항군집으로 구성되는 반면 선다형 검사의 경우 단순 지식 수준의 문항군집과 이해/적용 수준의 문항군집으로 분류된다. 보다 상세한 문항군집의 특성을 비교하기 위하여 하위 5개 문항군집(C1-C5)을 설정할 수 있다. 변형 선다형 검사의 군집 1은 국민 기본권에 관련된 문항, 군집 2는 문항형식에서 모두 고르기, 군집 3의 국민의 의무와 관련된 문항, 군집 4는 시사적 지식을 요구하는 문항, 군집 5는 법률 및 국가기관에 대한 지식을 묻는 문항으로 구성되었다. 이에 반하여 선다형 검사의 경우 군집 1은 국민 권리와 의무에 관한 문항, 군집 2는 지리관련 문항, 군집 3은 국가 기관 및 권리에 관한 문항, 군집 4는 의무에 관한 문항, 군집 5는 적용 수준의 문항으로 구성된다.

7) 군집의 수준을 상세화시키면 최종적으로 개별 문항 단위 군집을 설정할 수 있을 것이며, 이는 검사의 차원성을 해석하는 데 무의한 결과이다. 여기서 제시된 2개(T1, T2) 혹은 5개(C1-C5) 문항군집은 측정 구인에 대한 의미있는 해석을 위하여 문항내용, 형식, 지식수준에 기반하여 연구자가 설정한 것임을 밝힌다.



(a) 변형선다형 (b) 선다형

〔그림 2〕 군집 분석 결과

이러한 문항 군집 구조의 집단간 차별성은 다음 문항에 대한 피험자 반응 비교에서 전형적으로 나타난다.

17. 다음 중 국민의 기본적인 권리를 보장하기 위해 노력하는 국가 기관을 모두 고르시오.

① 청와대	② 국세청	③ 법률구조공단	④ 국가인권위원회

ㄱ. ①②③④ ㄴ. ①②③ ㄷ. ①②④ ㄹ. ①③④

위 문항은 17번 고유문항으로 '모두 고르시오.'라는 형식의 상대적으로 복잡한 사고 과정을 요구하는 문항으로 출제되었다. 실제 변형 선다형 집단에서 문항 정답률은 0.38이었으며, 문항특성이 〔그림 2〕에서 적용 수준(T2) 혹은 법률 및 국가조직 관련 문항(C5)으로 나타났다. 이에 반하여 선다형 집단에서 동일 문항의 정답률은 0.72로 상대적으로 높게 나타났으며, 단순 지식 수준(T1) 혹은 국가 기관 및 권리 관련 문항 군집(C3)에 속한다.

결국 동일한 문항임에도 불구하고 선택지를 어떤 형식으로 제시하느냐에 따라 문항 특성

및 검사의 차원성은 두 집단 사이에 일정한 차별성을 보인다. 특히 구성형과 선다형의 문항 특성을 포괄하는 변형선다형 검사는 전통적 선다형 검사와 비교하여 상위 인지 수준의 문항군집(2개 군집에서 적용 수준 문항군집, 5개 군집에서 모두 고르기 군집과 시사적 지식을 요구하는 문항군집)의 특성을 보다 상세히 변별하는 것으로 나타났다.

Ⅵ. 결론 및 논의

이 논문에서는 구성형 문항과 선다형 문항의 특성을 포괄하는 변형 선다형 검사를 소개하였으며, 변형 선다형 검사의 측정이론적 특성을 분석하기 위하여 구인의 차원성을 전통적 선다형 검사와 비교하였다. 구체적인 분석에 있어, 두 검사 형식의 분석을 위하여 다차원 문항반응모형과 군집 분석 절차가 활용되었다.

먼저 변형 선다형 집단과 선다형 집단의 능력 동등성을 확인하기 위하여 선다형 공통문항의 문항 정답률과 총점을 비교하였으며, 두 집단 사이에 통계적으로 유의미한 차이가 없는 것으로 나타났다. 부가적으로 검사점수 신뢰도를 비교한 결과 변형선다형 검사의 신뢰도가 선다형 검사의 신뢰도 보다 약간 높게 나타났다.

측정 구인의 차원성 분석을 위하여 6차원 문항반응모형이 적용된 결과로써 문항변별도에 있어 두 검사가 상당한 차이를 나타냈다. 즉 동일한 문항임에도 불구하고 문항 형식에 따라 주요하게 반응하는 차원이 개별 문항에 따라 다르게 나타났다. 이러한 차원성의 차이를 전체 검사 수준에서 비교하기 위하여 다차원 문항변별도로 계산된 각거리를 이용한 위계적 군집 분석을 적용하였다. 군집 분석 결과로 제시된 문항 군집의 경우, 내용영역이 군집내 일정 정도 혼재된 형태로 나타남에도 불구하고 변형 선다형 문항으로 구성된 검사가 상위 인지 수준의 문항 군집을 보다 상세히 변별하는 특징을 보인다.

이와 같은 변형선다형 검사의 문항군집 구조는 주요하게 문항 추측도와 관련된 것이라 할 수 있다. 다양한 선행 연구(Sireci & Zenisky, 2006; Wainer & Thissen, 1993; Zenisky & Sireci, 2002)에서 논의된 바와 같이, 선다형 문항에 대한 주요 비판점은 피험자가 정답을 정확히 알지 못하더라도 제시된 선택지를 활용하여 부분적 추측(informed guessing) 혹은 정답의 역추적이 가능하다는 것이다. 이에 반하여 변형 선다형 문항은 최초 문항을 단답형 형식으로 풀게하고 선택지를 상대적으로 짧은 시간만 제시하기 때문에 부분적 추측이나 정답의 역추적 가능성을 상당 부분 제거할 수 있다는 특성을 보인다. 더욱이 문항 풀이과정에서의 높은 인출난이도는 이후 기억 혹은 학습의 효과를 높이는 결과로 이어진다(Bjork, 1994; Park, 2005, 2007).

검사 점수의 해석을 위하여 활용되는 대표적인 측정이론인 고전검사이론과 문항반응이론은 암묵적으로 혹은 명시적으로 검사 점수의 단일 차원성(unidimensionality)을 가정한다. 구체적으로, 고전검사이론은 문항난이도(정답률)와 문항변별도(상관계수)의 계산에서 전체 단일 점수를 활용한다는 점에서 암묵적으로 일차원성을 가정한다. 이에 반하여 문항반응이론은 지역독립성과 피험자 반응의 조건부 독립성이라는 명시적 가정을 통하여 일차원성을 규정한다. 그러나 실제 많은 교육/심리 검사는 단일한 피험자의 능력, 지식, 기술을 측정하기 보다는 상호 관련된 다양한 차원에 작용하는 것으로 알려져 왔다(Ackerman, 1994). 이러한 선행 연구의 주장뿐만 아니라 검사 점수가 제공하는 정보의 다양성과 적절한 활용을 위하여 대규모 검사의 경우 다양한 하위 점수를 제공하고 있는 것이 근래의 추세이다. 구체적인 사례로 국제 학업성취도 비교 연구인 PISA와 TIMSS는 각 과목별로 다양한 하위 점수를 제공하고 이를 통하여 국가 수준의 학업 성취도를 세밀하게 평가하고자 한다(OECD, 2007). 우리나라의 초3 진단평가의 경우, 단일한 총점뿐만 아니라 하위 차원의 점수를 활용하여 학생 개인 수준의 진단 정보뿐만 아니라 지역, 국가 수준의 세부 정보를 제공한다.

다양한 하위 점수가 존재함에도 불구하고 이를 하나의 총점으로 표현하는 것은 수리적으로 서로 다른 공간에 존재하는 정보를 하나의 차원으로 투영(projection)하는 것으로 이러한 투영과정을 통하여 원래의 정보가 일부 손실될 뿐만 아니라 상이한 차원에 존재하는 정보가 혼합된 최종 점수는 그 의미가 명확하지 않아 해석에 어려움을 나타낸다.

다차원 문항반응이론은 이러한 검사-피험자의 상호작용의 복잡성이라는 현실을 분석모형에 반영한 측정이론으로써 검사 점수가 산출하는 다양한 정보의 차원성을 분할하는 이론으로 활용될 수 있다. 또한 검사의 차원성이 3차원 이상 복잡한 형태를 갖는 경우 이 논문에서 제시된 바와 같이 문항변별도와 각거리를 이용한 위계적 군집 분석을 통하여 적절한 수준의 검사 차원성을 분할/확인한다.

특히 근래 컴퓨터 기술의 발달에 따라 많은 검사 프로그램에 있어 컴퓨터를 활용한 검사의 시행 및 채점이 이루지고 있으며, 컴퓨터를 이용하기 때문에 적용될 수 있는 새로운 문항형식이 시도되고 있다. Huff과 Sireci(2001)이 논의한 바와 같이, 컴퓨터를 활용한 새로운 문항형식을 개발한 기존 연구의 가장 큰 문제점은 새로 개발된 문항이 측정하는 바가 무엇이며 기존 형식과 어떤 다른 점이 있는가를 명확히 규명하지 못한다는 것이다. 즉 전통적 문항형식과 비교하여 컴퓨터 환경에서 구현된 새로운 문항 형식이 실제 신뢰로운 검사점수를 산출하고 있는가 혹은 측정하는 구인이 보다 타당한가 등에 대한 규명이 새로운 문항 형식의 적용과 개선을 위한 선행 조건으로써 지속적으로 연구되어야 할 것이다.

참 고 문 헌

민경석(2004). 문항반응이론에서의 다차원적 접근. **교육평가연구, 17**(1), 15-31.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 18,* 255-278.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster Analysis.* Beverly Hills, California: SAGE.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human being. In J. Metcalfe & A. P. Shimamura(Eds.) *Metacognition.* MIT Press.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633-642.

Fraser, C. (1986). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* [Computer program]. Armidale, New South Wales, Australia: Center for Behavioral Studies, The University of New England.

Hirsch, T. M., & Miller, T. R. (1991, June). *Evaluation of a multidimensional item response theory procedure for investigating test dimensionality.* Paper presented at the annual meeting of the Psychometric Society, New Brunswick, New Jersey.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20,* 16-25.

Li, Y. M., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensionality IRT linking. *Applied Psychological Measurement, 24,* 115-138.

Miller, T. R., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory. *Applied Measurement in Education, 5,* 193-211.

Milligan, G. W. & Cooper, M. C. (1985). An examination of procedures for determining the number of cluster in a data set. *Psychometrika, 50,* 159-179.

OECD (2007). *PISA 2006: Science Competencies for Tomorrow's World.* OECD/PISA.

Park, J. (2005). Learning in a new computerized testing system. *Journal of Educational Psychology, 97,* 436-443.

Park, J. (2007). A new delivery system for CAT. In D. J. Weiss(Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*(Retrieved Aug, 2007 from www.psych.imn.edu/psylabs/CATCentral)

Park, J., & Choi, B. (2008). Higher retention after a new take-home computerized test. *British Journal of Educational Technology, 39,* 538-547.

Rodrigues, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40,* 163-184.

Rodrigues, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24,* 3-13.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Reckase, M. D. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Reckase, M. D., & Hirsch, T. M. (1991, April). *Interpretation of number-correct scores when the true number of dimensions assessed by a test is greater than two.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35,* 1-30.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna(Eds.), *Handbook of Test Development.* Mahwah, New Jersey: Lawrence.

Wainer, H., & Thissen, D. (1993). Combining multiple choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103-118.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15,* 337-362.

• *논문 접수 : 2008년 8월 28일 / 수정본 접수 : 2008년 10월 5일 / 게재 승인 : 2008년 10월 15일*

ABSTRACT

Dimensionality of Computerized Modified Multiple-choice Test

Kyung-Seok Min(Assistant Professor, Sejong University)
Jooyong Park(Associate Professor, Sejong University)

Selected response items, that are widely used in classroom and standardized assessment, have been criticized in that it was difficult to measure high order thinking abilities using these items. On the other hand, constructed response items seem to be inefficient in terms of content coverage and rater-reliability. These days, some computerized test programs try to adopt various innovative items that might combine strengths of two item formats. For example, Park(2005) developed computerized modified multiple-choice testing system(CMMT), in which examinees have to generate their own answers after looking at the stem only and then select the final answer from the given alternatives. In order to compare the construct structures measured by computerized modified multiple-choice tests and traditional multiple-choice tests, two test forms, consisting of common items and unique items, were administered to 6th-grade students in Seoul. Multidimensional IRT models and hierarchical cluster analyses were conducted. And issues for the further studies were discussed.

Key words : computerized modified multiple-choice tests, constructed response items, selected response items, multidimensional IRT models, hierarchical cluster analysis