교육과정평가연구
The Journal of Curriculum and Evaluation
2012, Vol. 15, No. 3, pp. 231~250

다차원 문항반응모형의 총점 산출을 위한 기준선 합성방법과 최대 검사정보함수 방법의 비교

민 경 석(세종대학교 부교수)*

≪ 요 약 ≫

교육/심리 검사가 서로 관련성을 갖는 다양한 능력을 측정함에 따라 측정모형은 전통적인 일차원 문항반응모형에서 다차원 문항반응모형으로 확장되어왔다. 다차원 문항반응모형을 이용한 신뢰로운 하위점수(sub-scores)의 산출방법에 대한 다양한 논의가 이루어져온 반면, 하위점수를 종합한 총점(overall scores)의 의미와 특성에 대한 연구는 상대적으로 제한적이라 할 수 있다. 이 연구의 목적은 피험자와 문항간의 상호작용에서 복수의 잠재적 능력을 가정하는 두 가지 총점 산출방법(기준선 합성방법과 검사정보함수 방법)의 특성을 비교하는 데 있다. 구체적으로 모의 자료(차원구조를 가정한 문항모수)와 실제 대규모 검사자료의 분석을 통하여 총점 산출방법에 따른 하위점수와 총점과의 관계를 논의하였다. 연구결과로써, 모의 자료 분석에서 검사의 차원구조가 단순구조에 가까울수록 기준선 합성방법과 최대 검사정보 방법의 총점 간 차이가 크게 나타났다. 이는 단순구조에서 총점-하위점수 사이에 단일한 선형관계를 가정할 수 없음을 의미한다. 두 번째, 복합구조를 갖는 실제 검사 자료 분석에서 두 가지 총점 산출방법이 유사한 결과를 제공함에도 불구하고, 검사정보함수 방법의 총점은 피험자 수준에 따라 상대적으로 중요한 능력수준에 대한 보다 세밀한 정보를 제공하는 특성을 보였다.

주제어 : 다차원 문항반응이론, 기준선 합성, 검사정보함수, 단순구조, 복합구조

Ⅰ. 서론

교육/심리 분야에서 피험자의 특성을 측정하기 위하여 활용되는 검사는 다양한 하위영역 혹은 내용으로 구성된다. 예를 들어 우리나라 국가수준 학업성취도 평가는 학생들의 학업 성취수

* 제1저자 및 교신저자, minkyungseok@sejong.ac.kr

준을 진단하기 위하여 국어, 영어, 수학 등과 같은 교과목별 검사로 구성되며, 각 교과목 검사는 다시 해당 과목의 하위 내용영역으로 세분화된다(김성숙, 상경아, 이상아, 2009). 표준화 검사 혹은 대규모 검사에서 보고되는 하위점수는 피험자의 강점과 약점에 대한 상세한 정보를 제공한다는 측면에서 진단적, 상담적 의미를 갖는다. 또한 이러한 하위영역점수를 종합한 검사 전체의 총점은 최종적이고 행정적 의사결정을 위하여 실제적 중요성을 갖는다. 예를 들어 미국 대학입학시험인 ACT 검사는 크게 영어, 수학, 읽기, 과학, 쓰기 등 5개 영역(과목)으로 구성되며, 이 중 수학영역은 내용적으로 연산, 대수, 기하 등으로 세분화된다. ACT 검사의 결과 보고서는 각 과목의 척도점수(1~36점)와 이를 평균한 총점을 포함한다(ACT, 2007, 2008). 즉, 하위점수는 진단적, 상담적 활용이라는 측면에서 중요한 의미를 갖는 반면, 총점은 행정적 의사결정을 위한 교육적 정보를 제공한다.

교육/심리 검사가 단일한 특성을 측정하는 수준을 넘어서 검사군(test battery)과 같이 서로 관련성을 갖는 다양한 능력을 측정함에 따라, 피험자와 문항간의 상호작용을 분석하기 위한 측정모형은 전통적인 일차원 문항반응모형에서 다차원 문항반응모형으로 확장되어왔다(Reckase, 2009). 다차원 문항반응모형이 검사자료 분석에 활용됨에 따라 보다 신뢰로운 하위점수(sub-scores)의 산출방법에 대한 다양한 논의가 이루져왔다. 즉, 하위 차원간의 상관을 모형에 반영하여 부가적인 정보를 활용하는 다차원 문항반응모형의 하위점수는 고전검사이론이나 일차원 문항반응모형 보다 신뢰로운 하위점수를 산출한다(민경석, 2011; DeMars, 2005; Wang, Chen, & Cheng, 2004; Yao, 2010, 2011; Yao & Boughton, 2007). 그러나 고전검사이론의 총점에 대한 다양한 논의(김신영, 1999; Oosterhof, 1987; Sinharary, Haberman, & Puhan, 2008)와는 달리 다차원 문항반응모형으로부터 추정된 하위점수를 종합하여 전체 검사에 대한 점수를 산출하는 총점(overall scores)의 의미와 특성에 대한 연구는 상대적으로 제한적이라 할 수 있다(강태훈, 2010; Yao, 2010).

다차원 문항반응이론에서 잠재적 차원별로 산출된 능력 추정치를 합성하여 총점을 계산하는 방법으로 기준선 합성방법(reference composite method)과 검사정보함수 방법(test information function method) 등 두 가지가 제안되었다. 그러나 다차원 문항반응이론의 총점 산출모형에 대한 선행연구는 일차원 문항반응모형과의 일치성을 확인하는 수준에 머무르며(Ackerman, 1994; Wang, 1985, 1986; Yao, 2010, 2011; Yao & Schwarz, 2006), 두 가지 총점 산출방법의 직접적인 비교는 논의되지 못하는 제한점을 보인다. 특히, 다양한 검사 특성을 고려할 때, 차원구조(단순구조와 복합구조)는 다차원 문항반응모형의 하위 능력모수(하위점수)의 신뢰도에 영향을 미치며(민경석, 2011; Boughton, Yao, & Lewiset, 2006; Yao, 2010), 이러한 하위점수를 합성하는 다차원 모형의 두 가지 총점 산출방법은 차원구조에 따라 측정이론적 적절성이 판단될 수 있을 것이다.

이 연구의 목적은 피험자와 문항간의 상호작용에서 복수의 잠재적 차원을 가정하는 두 가지

총점 산출방법(기준선 합성방법과 검사정보함수 방법)의 특성을 논의하는 데 있다. 구체적으로 모의 자료를 활용하여 검사의 차원구조에 따른 기준선 합성방법과 검사정보함수 방법의 유사성과 차별성을 비교하였다. 또한 실제 대규모 검사자료의 분석을 통하여 총점 산출방법에 따른 각 차원의 하위점수와 총점과의 관계를 논의하였다.

이상의 연구 목적에 기반하여 2장에서는 다차원 문항반응이론과 총점 산출방법의 수리 모형을 제시했으며, 3장에서는 검사의 차원성을 가정한 가상적 문항모수와 49개 능력수준 자료를 분석하여 두 가지 총점 산출방법을 비교했으며, 4장에서는 실제 시행된 검사자료 분석을 통하여 하위점수와 총점간의 관계를 논의했다. 마지막으로 5장에서 연구결과에 대한 논의와 후속 연구를 위한 제언을 제시했다.

Ⅱ. 다차원 문항반응이론과 총점 산출모형

1. 다차원 문항반응모형

일차원 문항반응모형과 비교하여 다차원 문항반응모형은 문항과 피험자의 상호작용의 복잡성을 측정모형에 반영한 것이라 할 수 있다. 다양한 다차원 문항반응모형 중 가장 널리 활용되는 Reckase(2009)의 보완 다차원 문항반응모형(compensate multidimensional IRT models)의 문항 정답 확률은 식 (1)로 정의된다.

$$P(u_{ij}=1|\mathbf{a}_i, c_i, d_i; \boldsymbol{\theta}_j) = c_i + (1-c_i)\frac{\exp(\mathbf{a'_i\theta_j} + \mathrm{d_i})}{1+\exp(\mathbf{a'_i\theta_j} + \mathrm{d_i})} \quad (1)$$

식 (1)에서 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($m$은 능력차원 수), $\mathbf{a}_i = (a_{i1}, a_{i2}, \cdots, a_{im})'$는 i번째 문항의 변별도 벡터, d_i는 i번째 문항의 난이도관련 위치모수, c_i는 i번째 문항의 추측도관련 모수를 각각 의미한다.

다차원 문항반응모형은 식 (1)에서 제시된 바와 같이 둘 이상의 잠재적 차원에 대한 모수 추정치를 제공한다는 측면에서 일차원 문항반응모형과 구분된다. 즉, 실제 검사자료가 다차원적인 경우, 일차원 문항반응모형은 여러 차원의 능력모수를 하나의 능력 추정치로 투영(projection)하는 반면, 다차원 문항반응모형은 각각의 차원에 대하여 구분되는 정보를 제공한다는 장점을 보인다.

이와 같이 다차원 문항반응모형은 이론적으로 가정되는 혹은 통계적으로 설정된 여러 개 차

원의 능력을 추정하는 측정모형이며, 이를 통하여 총점(전체 점수)를 계산하는 것은 논리적으로 적절하지 못하다. 즉, 검사의 단일 총점은 명시적 혹은 비명시적으로 일차원성을 가정하는 것으로 이는 다차원 모형의 기본 가정과 배치된다(Reckase, 2009). 그럼에도 불구하고, 검사에서 산출되는 하위점수가 서로 독립적이기 보다는 일정정도의 상관를 보이며, 총점을 통한 행정적 의사결정을 위하여 다차원 문항반응이론의 총점 산출모형이 기준선 합성방법(Reckase, 2009; Wang, 1985, 1986)과 검사정보함수 방법(Ackerman, 1994; Yao, 2011; Yao & Schwarz, 2006)으로 제안되었다.

2. 기준선 합성방법

기준선 합성방법(reference composite method)은 다차원 문항반응모형에서 추정된 2개 이상의 능력 추정치를 선형적으로 합성하는 방법이다(Reckase, 2009; Wang, 1985, 1986). 즉, 기준선 합성방법은 전통적인 요인분석 절차와 유사하게, 여러 변수(하위점수)의 공통 분산에 근거하여 설명력이 가장 높은 하나의 공통 요인(총점)을 추출하는 방법이라 할 수 있다.

구체적으로, 기준선 합성방법은 검사에 포함된 문항의 변별도가 평균적으로 가장 높아지는 방향으로 문항점수를 합성하며, 하위점수의 가중치는 문항 변별도 행렬에 대한 고유값-고유벡터 분할(eigen value-eigen vector decomposition)을 통하여 결정된다. 즉, $\mathbf{A}=(\mathbf{a_1}, \mathbf{a_2}, \cdots, \mathbf{a_i})'$라고 할 때 교차행렬 A′A은 식 (2)와 같이 고유값-고유벡터로 분할된다.

$$(A'A-\lambda\mathbf{I})\mathbf{w}=0 \qquad (2)$$

식 (2)에서 λ는 교차행렬의 고유값, $\mathbf{I}$는 단위행렬, $\mathbf{w}$는 고유값에 대응하는 고유벡터를 나타낸다.

일반적으로 m차원으로 구성된 검사의 문항 변별도 교차행렬은 m개의 고유값을 가지며, 이 중 가장 큰 고유값에 대응하는 고유벡터를 이용하여 기준선 합성방법은 식 (3)과 같이 하위점수를 합성하여 총점을 산출한다.

$$\theta = w_1\theta_1 + w_2\theta_2 + \ldots + w_m\theta_m, \text{ 그리고 } \sum_{i=1}^{m} w_i^2 = 1 \qquad (3)$$

식 (3)은 하위능력을 선형적으로 조합하여 총점(θ)을 추정하는 것이며, 이때 각 차원의 능력($\theta_1, \theta_2, \cdots, \theta_m$)에 대한 가중치로 식 (2)에서 계산된 최대 고유값에 대응하는 고유벡터의 요소를 이용한다. 결국, 기준선 합성방법은 모든 능력수준에서 하위점수와 총점간의 동일한 선형관계를 가정하며, 검사를 구성하는 문항 변별도가 평균적으로 최대가 되는 방향으로 각 차원에 대한 가중치를 설정한다.

3. 검사정보함수 방법

검사정보함수 방법은 검사정보가 최대화되는 방향으로 총점을 산출하는 것으로, 다차원 문항반응모형의 문항정보함수는 식 (4)와 같다(Yao & Schwarz, 2006).

$$I_i(\boldsymbol{\theta}) = \left(\frac{P_i(\boldsymbol{\theta}) - c_i}{1 - c_i} \right)^2 \frac{Q_i(\boldsymbol{\theta})}{P_i(\boldsymbol{\theta})} (\mathbf{a}_i \otimes \mathbf{a}_i) \tag{4}$$

식 (4)에서 $Q_i = (1 - P_i)$이며, 행렬 연산기호 $\otimes$는 외적(outer product)을 의미한다. 식 (4)의 문항정보함수를 모든 문항에 대하여 합산하면 검사정보함수가 구해지며, 이에 대하여 역수를 취하면 식 (5)와 같이 최대우도추정방법(maximum likelihood estimation)에 의한 능력모수 추정치의 근사적 오차분산이 계산된다.

$$V(\boldsymbol{\theta}) = [I(\boldsymbol{\theta})]^{-1} = [\sum I_i(\boldsymbol{\theta})]^{-1} \tag{5}$$

식 (5)에서 $V(\boldsymbol{\theta})$은 특정 능력수준($\boldsymbol{\theta}$)에서의 오차분산을 의미하며, 다차원 문항반응모형의 오차분산은 측정 방향의 미결정성(directional indeterminacy, Ackerman, 1994; Reckase, 2009)때문에 식 (6)과같이 벡터공간의 특정 방향(α)에 따라 다른 값을 갖는다.

$$V(\boldsymbol{\theta}_\alpha) = \boldsymbol{w}' V(\boldsymbol{\theta}) \boldsymbol{w}, \text{ 그리고 } \boldsymbol{w}'\boldsymbol{w} = 1 \tag{6}$$

식 (6)에서 $V(\boldsymbol{\theta}_\alpha)$는 특정 방향($\alpha$)에 대한 오차분산, $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)'$는 방향성을 결정하는 가중치를 나타낸다. 그러므로 식 (6)의 오차분산을 최소화하는 가중치 벡터 $\boldsymbol{w}$를 통하여 최대 검사정보를 갖는 총점이 찾아진다.

기준선 합성방법과 검사정보함수 방법의 수리적 차별성은 식 (2)와 (6)에 포함된 가중치($\mathbf{w}$)에서 주요하게 나타난다. 즉, 식 (2)에서 계산된 가중치는 모든 피험자에게 동일하게 적용되며, 이는 하위점수와 총점간에 일관된 선형관계를 의미한다. 이에 반하여 식 (6)에서 계산되는 가중치는 피험자의 능력수준에 따라 다른 값을 가지며, 하위점수와 총점간의 선형관계는 피험자의 위치에 따라 변화하게 된다.

Ⅲ. 모의 자료 분석

1. 검사의 차원구조

하위점수의 수리적 합성인 총점에 영향을 주는 검사의 차원구조는 일반적으로 단순구조와 복합구조로 구분된다. 이론적으로 순수한 단순구조(문항 변별도가 특정 차원에서 높고, 나머지 차원에서는 모두 0)는 개별 문항이 하나의 잠재적 차원에만 작용하여 측정되는 구인의 특성이 문항별로 명확히 구분되는 상황이라 할 수 있다. 그러나 교육/심리 검사에서 측정되는 잠재적 특성은 서로 연관성을 가지며, 때론 하나의 구인의 서로 다른 측면으로 구성되기 때문에 유사 단순구조(문항 변별도가 특정 차원에서 높고, 나머지 차원에서는 0에 가까움)가 보다 현실적인 가정이라 할 수 있다. 단순구조와 비교되는 복합구조는 측정구인이 이론적으로 서로 구분되어짐에도 불구하고 개별 문항의 변별도가 두 개 이상의 차원에서 0과 다른 값은 갖는 경우라 할 수 있다.

검사의 차원구조에 따른 다차원문항반응모형의 총점 산출방법의 비교를 위하여 2차원의 20개 문항으로 구성된 검사를 가정하였으며, 능력모수는 -3~+3 범위에서 두 차원이 교차하는 49개 능력수준을 비교하였다.

먼저 유사 단순구조와 복합구조를 갖는 20개의 가상적 문항의 모수를 제시하면 〈표 1〉과 같다.[1)]

〈표 1〉 20개 문항의 모수

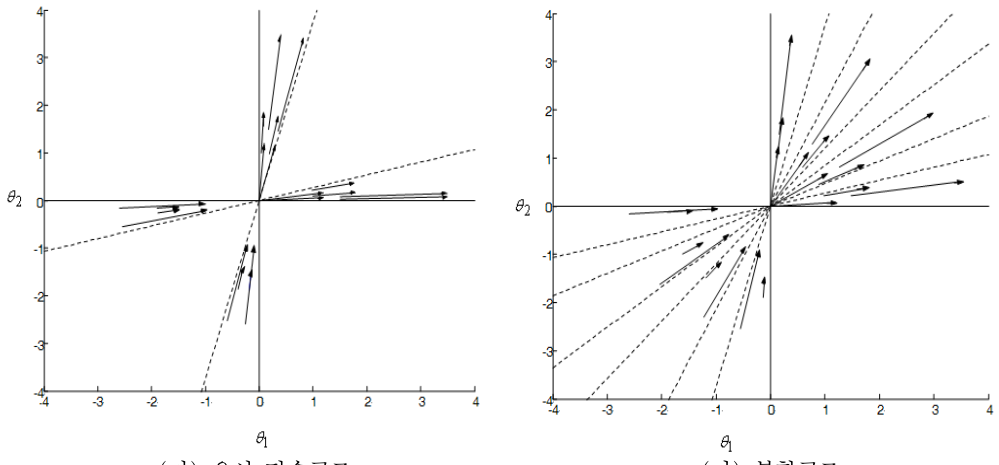
문항번호	유사 단순구조		복합구조		d
	a1	a2	a1	a2	
1	***0.40***	0.03	***0.40***	0.03	0.60
2	***0.80***	0.07	***0.78***	0.17	0.80
3	***1.19***	0.16	***1.20***	0.07	0.00
4	***1.56***	0.34	***1.60***	0.10	1.60
5	***2.00***	0.04	***1.98***	0.29	3.00
6	***0.40***	0.05	***0.34***	0.21	0.60
7	***0.78***	0.17	***0.71***	0.36	0.80
8	***1.20***	0.06	***1.01***	0.64	0.00
9	***1.60***	0.11	***1.25***	1.00	1.60
10	***2.00***	0.09	***1.68***	1.08	3.00
11	0.04	***0.40***	0.25	***0.31***	0.60
12	0.15	***0.79***	0.47	***0.65***	0.80
13	0.09	***1.20***	0.64	***1.01***	0.00

1) 유사 단순구조와 복합구조 모두에서 20개 문항의 추측도는 0으로 고정하였다.

문항번호	유사 단순구조		복합구조		d
	a1	a2	a1	a2	
14	0.16	***0.59***	0.75	***1.41***	1.60
15	0.47	***1.94***	1.03	***1.71***	3.00
16	0.08	***0.39***	0.03	***0.40***	0.60
17	0.04	***0.80***	0.10	***0.79***	0.80
18	0.30	***1.16***	0.14	***1.19***	0.00
19	0.37	***1.56***	0.34	***1.56***	1.60
20	0.23	***1.99***	0.21	***1.99***	3.00
평균	0.69	0.65	0.75	0.75	0.32
표준편차	0.66	0.69	0.57	0.60	1.59

〈표 1〉에 제시된 문항모수는 Roussos 외(1998)가 이용한 5개 문항의 다차원 문항 변별도와 문항 난이도를 기준으로 반복적으로 생성된 것이다. 유사 단순구조에서 20개 문항은 두 개의 문항군으로 구성되며, 이중 중 10개 문항은 첫 번째 차원에 주로 작용하고, 나머지 10개 문항은 두 번째 차원에 집중적으로 부하되는 차원구조를 보인다. 이에 반하여 복합구조에서는 검사문항이 4개 군집(5개 문항)을 이루며, 2개 군집은 두 차원 중 하나에만 주로 작용하고, 나머지 두 개 문항군집은 모든 차원에 고르게 반응하는 특성을 보인다.

〈표 1〉의 차원구조에 따른 문항모수를 좌표 공간의 문항벡터(item vectors)[2]로 표현하면 〔그림 1〕과 같다.



(가) 유사 단순구조 (나) 복합구조

〔그림 1〕 20개 문항의 차원구조

2) 문항 벡터는 난이도(벡터와 원점과의 거리), 변별도(벡터의 길이), 최대문항정보 방향(문항벡터와 축이 이루는 각도) 등의 세 가지 문항 정보를 기하학적으로 표현한 것이다(Reckase, 2009).

〔그림 1〕에 제시된 바와 같이 유사 단순구조에서 두 개 문항군은 각 좌표축에 가깝게 위치한다(문항벡터가 좌표축과 15° 이내의 각도에 위치함). 이에 반하여 복합구조에서 2개 문항 군집이 좌표축과 이루는 각도가 15° 이내이며 나머지 두 개 문항군집은 25°~40° 수준이다.

2. 총점의 방향

〈표 1〉에 제시된 문항 변별도에 식 (2)와 (3)을 적용하여 기준선 합성방법의 총점 기준선을 결정하면, 유사 단순구조의 최대 고유값에 대응하는 고유벡터는 $(0.74,\ 0.68)'$이다. 이에 따라 총점은 $\theta = 0.74 \times \theta_1 + 0.68 \times \theta_2$으로 계산된다. 이때 기준선이 좌표평면의 두 축과 이루는 각도는 (42.6°, 47.4°)이며[3], 총점 선형식에서 첫 번째 가중치가 상대적으로 크기 때문에 첫 번째 축과 이루는 각도가 작게 나타난다(부하량이 높음). 복합구조에서 최대 고유값에 대응하는 고유벡터는 $(0.69,\ 0.72)'$이며(즉, $\theta = 0.69 \times \theta_1 + 0.72 \times \theta_2$), 복합구조의 기준선이 두 축과 이루는 각은 (46.0°, 44.0°)과 같다. 〔그림 1〕에 나타난 바와 같이 유사 단순구조와 복합구조의 20개 문항이 전체적으로 첫 번째 차원과 두 번째 차원에 대칭적으로 분포함에 따라, 기준선 합성방법의 총점 방향은 균일 가중치를 의미하는 45°선에 가깝게 나타났다. 결국, 기준선 합성방법은 개별 문항이 측정하는 방향이 서로 다름에도 불구하고, 유사 단순구조와 복합구조 모두에서 평균적으로 두 차원에 균일한 가중치가 부여되는 총점이 산출된다.

이에 반하여 식 (4)-(6)을 이용한 최대 검사정보함수 총점의 방향은 능력수준에 따라 다르게 나타난다. 이를 확인하기 위하여, 49개 능력수준에서 첫 번째 차원에 대한 검사정보함수 총점 방향을 제시하면 〈표 2〉와 〈표 3〉과 같다.

〈표 2〉 유사 단순구조에서 검사정보함수 방법의 총점 방향(첫 번째 차원과의 각도)

θ_2								
	3	24.17	6.55	6.66	12.63	16.78	30.73	67.71
	2	72.50	28.16	24.56	48.13	51.19	65.84	76.27
	1	76.87	58.71	44.84	55.05	51.17	66.03	76.70
	0	75.36	42.85	29.90	43.26	45.00	67.09	77.37
	-1	74.54	39.60	29.96	40.91	43.14	67.02	77.54
	-2	69.96	20.45	15.08	18.89	19.66	41.04	72.03
	-3	34.71	7.47	7.23	9.70	10.20	14.31	39.70
		-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00
					θ_1			

3) 가중치와 각도는 삼각함수를 통하여 수리적으로 동일한 의미를 갖는다. 예를 들어, 유사 단순구조 가중치 (0.74, 0.68)에 아크코사인(arc cosine)을 취하면 $\arccos(0.74,\ 0.68) \times 180/\pi = (42.6°,\ 47.4°)$이다.

〈표 3〉 복합구조에서 검사정보함수 방법의 총점 방향(첫 번째 차원과의 각도)

	3	45.80	25.16	22.50	30.08	35.48	45.25	51.68
	2	66.69	45.92	38.40	44.42	48.10	54.28	60.51
	1	67.09	51.43	44.92	48.25	48.70	54.91	62.62
θ_2	0	57.47	44.52	41.83	45.68	47.22	55.80	64.97
	-1	51.97	42.43	43.82	46.26	46.42	55.50	65.64
	-2	47.78	40.25	40.32	38.52	33.91	41.99	57.65
	-3	42.00	32.42	31.57	25.96	18.71	22.60	38.66
		-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00
					θ_1			

〈표 2〉와 〈표 3〉은 최대 검사정보를 갖은 총점이 첫 번째 차원(θ_1)과 이루는 각도를 나타낸다.[4] 예를 들어, 유사 단순구조에서 피험자 능력수준이 ($\theta_1=0$, $\theta_2=0$) 경우, 검사정보함수의 방향은 (43.26°, 〔90-43.26〕°)으로 이는 기준선 합성방법의 방향(42.6°, 47.4°)과 유사하다. 그러나 또 다른 능력수준인 ($\theta_1=-3$, $\theta_2=3$)에서 검사정보방법의 총점 방향은 (24.17°, 65.83°)으로 기준선 방향과 매우 다르게 첫 번째 차원의 부하량이 높아진다.

모든 피험자에게 동일한 가중치가 부여되는 기준선 합성방법과 달리, 〈표 2〉와 〈표 3〉에 나타난 바와 같이 검사정보함수의 총점 방향(가중치)은 피험자의 능력수준에 따라 다르게 설정된다. 구체적으로, 유사 단순구조에서 중간 능력수준(〈표 2〉의 음영 부분)은 각 차원에 대하여 45°(기준선 합성방법과 유사한 방향)에 상대적으로 가까운 반면, 각 차원의 능력수준이 극단값(-3 혹은 +3)에 가까울수록 다른 차원의 영향이 커지는 것으로 나타났다. 예를 들어 유사 단순구조인 〈표 2〉에서 첫 번째 차원의 능력이 -3, 혹은 +3으로 매우 낮거나 높을 때 하위점수의 가중치 값은 두 번째 차원에서 높아지는 경향을 보인다(첫 번째 차원과의 각이 커지고 두 번째 차원과 이루는 각이 작아짐). 이러한 경향은 피험자가 두 번째 차원에서 극단의 능력수준일 때 첫 번째 차원의 가중치가 높아지는 것으로 동일하게 나타난다.

피험자의 능력수준에 따른 가중치의 변화 경향은 〈표 3〉에 제시된 복합구조에서도 동일하게 나타난다. 그러나 복합구조는 유사 단순구조와 달리 기준선과 유사한 방향을 갖는 중간능력 수준(음영 부분)의 범위가 상대적으로 넓으며, 각 차원에서 능력수준이 극단값을 가질지라도 다른 차원의 가중치가 상대적으로 적게 커진다.

결국, 검사정보함수 방법은 오차분산을 최소화하는 총점을 산출하는 측정이론적 장점을 가짐에도 불구하고 피험자에 따라 가중치가 달라지며, 이러한 특성은 복합구조 보다 단순구조에서 강하게 나타난다.

4) 좌표평면에서 두 번째 차원(θ_2)과 이루는 각은 90°에서 표의 값을 차감함으로써 계산된다.

3. 총점간 상관

총점 산출방법에 따라 총점의 방향이 달라진 결과를 보다 구체적으로 확인하기 위하여, 유사 단순구조와 복합구조의 49개 능력수준에 대한 두 가지 총점과 상관을 비교하면 〈표 4〉와 〈표 5〉와 같다.

〈표 4〉에 제시된 바와 같이, 유사 단순구조와 복합구조 모두에서 기준선 합성방법은 두 능력 차원에 유사한 가중치(45° 방향)를 부여함에 따라 큰 차이를 보이지 않고 있다. 그러나 각 차원 구조 내에서 기준선 합성 총점과 검사정보함수 총점은 일정한 차이가 나타난다. 특히 유사 단순구조에서 두 가지 총점의 차이가 매우 큰 경우(1.0 이상)가 자주 발견된다(음영 부분). 이에 반하여 복합구조에서 두 가지 총점은 그 차이가 상대적으로 작게 나타난다.

〈표 4〉 차원구조에 따른 총점 결과

능력모수		단순구조		복합구조		능력모수		단순구조		복합구조	
θ_1	θ_2	기준선	검사정보	기준선	검사정보	θ_1	θ_2	기준선	검사정보	기준선	검사정보
-3	-3	-4.24	-3.97	-4.24	-4.24	0	2	1.35	0.65	1.44	1.25
-3	-2	-3.56	-2.81	-3.52	-3.02	0	3	2.03	0.51	2.16	1.31
-3	-1	-2.89	-1.66	-2.8	-2.09	1	-3	-1.29	0.09	-1.47	-0.93
-3	0	-2.21	-0.76	-2.08	-1.61	1	-2	-0.62	-0.93	-0.75	-0.82
-3	1	-1.53	0.16	-1.36	-1.06	1	-1	0.06	-0.15	-0.03	-0.09
-3	2	-0.86	0.85	-0.64	-0.53	1	0	0.74	0.71	0.69	0.68
-3	3	-0.18	-0.76	0.08	-0.22	1	1	1.41	1.41	1.41	1.41
-2	-3	-3.5	-2.33	-3.55	-3.09	1	2	2.09	1.61	2.13	1.95
-2	-2	-2.83	-2.71	-2.83	-2.83	1	3	2.77	1.52	2.85	1.91
-2	-1	-2.15	-1.89	-2.11	-2.03	2	-3	-0.56	0.19	-0.77	-0.72
-2	0	-1.47	-1.47	-1.39	-1.43	2	-2	0.12	-1.01	-0.05	-0.46
-2	1	-0.8	-0.9	-0.67	-0.8	2	-1	0.8	-0.1	0.67	0.33
-2	2	-0.12	-1.18	0.05	-0.23	2	0	1.47	0.78	1.39	1.12
-2	3	0.56	-1.59	0.77	-0.08	2	1	2.15	1.7	2.11	1.96
-1	-3	-2.77	-1.34	-2.85	-2.07	2	2	2.83	2.82	2.83	2.82
-1	-2	-2.09	-1.74	-2.13	-2.03	2	3	3.5	2.68	3.55	3
-1	-1	-1.41	-1.41	-1.41	-1.41	3	-3	0.18	-1.64	-0.08	-0.49
-1	0	-0.74	-0.87	-0.69	-0.75	3	-2	0.86	-1.23	0.64	-0.26
-1	1	-0.06	-0.37	0.03	-0.03	3	-1	1.53	-0.28	1.36	0.49
-1	2	0.62	-0.45	0.75	0.53	3	0	2.21	0.66	2.08	1.27
-1	3	1.29	-0.61	1.47	0.72	3	1	2.89	1.62	2.8	2.15
0	-3	-2.03	-0.66	-2.16	-1.5	3	2	3.56	2.83	3.52	3.29

능력모수		단순구조		복합구조		능력모수		단순구조		복합구조	
θ_1	θ_2	기준선	검사정보	기준선	검사정보	θ_1	θ_2	기준선	검사정보	기준선	검사정보
0	-2	-1.35	-1.49	-1.44	-1.4	3	3	4.24	4.22	4.24	4.22
0	-1	-0.68	-0.82	-0.72	-0.75	평균		0.00	-0.23	0.00	-0.12
0	0	0	0	0	0	표준편차		2.02	1.60	2.02	1.76
0	1	0.68	0.65	0.72	0.72						

〈표 5〉 차원구조에 따른 총점 상관

구분		유사 단순구조		복합구조	
		기준선	검사정보	기준선	검사정보
유사 단순구조	기준선	1	.81	-	-
	검사정보	.87	1	-	-
복합구조	기준선	-	-	1	.99
	검사정보	-	-	.98	1

차원구조에 따른 이러한 차이는 총점간의 상관을 제시한 〈표 5〉에서도 확인된다. 〈표 5〉에서 대각선을 기준으로 아래는 단순상관(Pearson correlation), 위는 서열상관(Spearman rank correlation)을 나타낸다. 〈표 4〉에서 논의된 바와 같이 유사 단순구조에서 검사정보함수 방법과 기준선 합성방법의 상관은 복합구조 보다 낮게 나타난다.

요약하면, 문항에 따라 측정구인이 상대적으로 명확히 구분되는 단순구조에서 검사정보함수 방법의 하위영역별 가중치가 피험자 수준에 따라 크게 달라지며, 이에 따라 기준선 합성방법과 검사정보함수 방법의 총점간 상관이 낮아지는 경향을 보인다.

Ⅳ. 실제 검사자료 분석

1. 실제 검사자료

모의 자료 분석에서는 검사의 차원구조와 피험자의 능력수준이 알려졌음을 가정한다. 그러나 현실적으로 연구자는 실제 검사자료의 차원구조를 알지 못하며, 출제 계획서를 통하여 검사의 차원구조를 이론적으로 가정하거나, 요인분석과 같은 경험적 방법을 통하여 차원구조를 추정한다. 이러한 검사자료 분석의 현실적 상황을 고려하여 다차원 문항반응모형의 두 가지 총점 산출

방법을 비교하기 위하여 실제 시행된 대규모 검사자료를 분석하였다. 실제 자료 분석을 위하여 외국인의 한국어 능력을 측정하는 한국어능력시험의 하위 검사인 어휘/문법 검사를 이용하였다. 한국어능력시험의 어휘/문법 검사는 총 30문항(4지 선다형 문항)으로 구성되며, 하위내용인 어휘 영역과 문법 영역이 각 15개 문항으로 출제된다. 2011년 실시된 초급 어휘/문법 검사에 총 6,560명이 응시했으며, 이중에서 임의추출된(random sampling) 2,000명의 자료를 분석에 활용했다. 표집된 어휘/문법 검사자료의 원점수 기술 통계치는 〈표 6〉과 같다.

〈표 6〉 어휘/문법 검사의 기술 통계치 (원점수)

영역	사례수	평균	표준편차	분산	최소	최대
어휘	2,000	12.15	2.72	7.37	1.00	15.00
문법		10.72	2.98	8.90	.00	15.00
총점		22.86	5.34	28.47	2.00	30.00

어휘/문법 검사의 하위영역인 어휘와 문법 점수의 상관은 0.77이며, 각 하위영역의 총점간 상관은 0.93, 0.95이다. 대규모 검사에서 측정되는 하위영역의 상관이 일반적으로 0.3~0.8 정도인 점을 고려할 때(Yao, 2010), 어휘/문법 검사의 두 영역간 상관은 상대적으로 높은 수준이라 할 수 있다.

2. 하위점수 추정과 총점 산출

어휘/문법 검사에 대하여 내용적 측면(출제 계획서에 근거함)에서 2개의 잠재적 차원(어휘, 문법)이 설정됨에 따라, 2차원 3모수 문항반응모형을 적용하였다. 문항모수와 능력모수를 추정하기 위하여 TESTFACT(Bock, et al., 2003)를 이용하였으며, TESTFACT 프로그램이 추측도를 제공하지 않기 때문에 BILOG-MG(Zimowski, et al., 1996)로 추정된 추측도를 입력자료로 활용하였다. 또한 각 차원의 능력모수는 사후분포기대치(expected a posteria estimates)로 추정되었다. 2차원 3모수 문항반응모형으로 추정된 30문항의 모수 추정치는 〈표 7〉과 같다.

〈표 7〉 어휘/문법 검사의 문항모수 추정치

문항번호	언어 영역 변별도(a1)	문법 영역 변별도(a2)	난이도(d)	추측도(c)
1	*1.06*	0.10	3.00	0.25
2	*1.11*	0.30	2.42	0.21
3	*2.26*	1.01	3.73	0.18

문항번호	언어 영역 변별도(a1)	문법 영역 변별도(a2)	난이도(d)	추측도(c)
4	***0.76***	0.63	1.16	0.20
5	***1.01***	0.88	1.59	0.25
6	***2.78***	1.48	4.03	0.20
7	***1.21***	0.85	2.33	0.20
8	***1.60***	1.00	2.67	0.19
9	***1.11***	0.55	1.43	0.17
10	***0.63***	0.29	0.46	0.22
11	0.96	***1.27***	0.41	0.21
12	0.65	***0.83***	0.49	0.14
13	***0.69***	0.53	1.20	0.24
14	***0.69***	0.36	1.53	0.24
15	0.96	***1.66***	1.09	0.18
16	***0.52***	0.12	0.43	0.19
17	***1.06***	0.60	2.40	0.26
18	***1.31***	0.65	2.09	0.18
19	0.29	***0.99***	-0.63	0.22
20	***1.12***	0.51	1.30	0.15
21	0.23	***0.74***	0.04	0.28
22	0.75	***2.71***	0.61	0.20
23	***0.58***	0.43	0.10	0.12
24	0.62	***0.57***	-0.92	0.29
25	1.67	***2.65***	2.01	0.21
26	1.27	***2.40***	1.58	0.23
27	0.48	***0.67***	-0.12	0.15
28	0.46	***0.89***	-0.77	0.13
29	0.20	***1.44***	-0.15	0.30
30	0.60	***1.09***	0.50	0.28
평균	0.95	0.94	1.20	0.21
표준편차	0.57	0.68	1.27	0.05

어휘/문법 검사의 출제 계획서에 근거할 때, 어휘 영역을 측정하는 문항(문항번호 1~15)은 첫 번째 변별도가 상대적으로 높으며, 문법 영역 문항(문항번호 16~30)은 두 번째 변별도가 높을 것으로 기대된다. 〈표 7〉에 제시된 문항 변별도는 이러한 이론적 차원구조에 대체적으로 부합한다(각 문항에서 두 차원 중 상대적으로 높은 변별도가 굵게 표시됨.). 또한 대부분의 문항에서 부차적인 차원의 변별도가 0보다 높은 값을 나타낸다. 즉, 자료 분석에 활용된 어휘/문

법 검사는 복합구조의 차원구조를 보인다.

복합구조를 갖는 어휘/문법 검사의 총점을 구하기 위한 기준선 합성방법의 가중치(고유벡터)는 (0.69, 0.73)이며, 총점은 $\theta = 0.69 \times \theta_1 + 0.73 \times \theta_2$의 선형식으로 계산된다. 이에 따라 총점 기준선이 좌표평면에서 두 축(어휘 영역과 문법 영역)과 이루는 각도는 (46.86°, 43.51°)로 균일 가중치에 해당하는 45°선에 가깝게 나타났다. 이러한 결과는 〈표 7〉에 제시된 문항 변별도 평균(0.95, 0.94)이 두 영역에서 유사하다는 점이 반영된 것이다.

다차원 문항반응모형의 검사정보함수를 이용한 총점의 가중치는 능력수준에 따라 다르며, 이를 평균한 가중치는 (0.58, 0.80)이고, 평균 가중치를 각 축과 이루는 각도로 표현하면 (54.54°, 36.87°)[5]이 된다. 결국, 어휘/문법 검사의 기준선 합성방법은 두 차원의 가중치가 균등한 반면, 검사정보함수 방법의 가중치는 평균적으로 두 번째 차원에서 상대적으로 높게 나타났다.

2차원 3모수 문항반응모형으로 추정된 하위점수를 합성하는 두 가지 방법에 의한 총점의 기술 통계치와 상관은 〈표 8〉과 〈표 9〉와 같다. 〈표 8〉과 〈표 9〉에서 추가적으로 일차원 문항반응모형의 능력 추정치의 결과를 함께 제시했다.

〈표 8〉 어휘/문법 검사의 총점 기술 통계치

구분	사례수	평균	표준편차	최소	최대
일차원 모형	2,000	.00	.95	-2.88	1.73
기준선 합성		-.04	.89	-3.73	1.68
검사정보		-.01	.86	-2.86	1.61

〈표 9〉 어휘/문법 검사의 총점간 상관

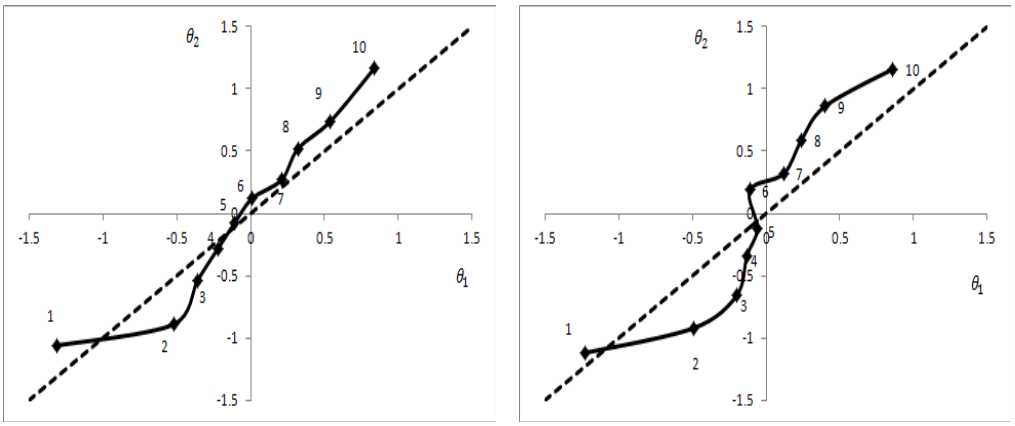
구분	일차원 모형	기준선 합성	검사정보
일차원 모형	1	.99	.99
기준선 합성	.99	1	.99
검사정보	.98	.99	1

〈표 8〉에 제시된 바와 같이, 일차원 문항반응모형의 총점과 비교하여 다차원 문항반응모형의 두 가지 총점은 유사한 평균 점수를 보이는 반면, 표준편차가 약간 작아지는 경향을 보였다. 즉, 다차원 모형의 총점은 상관을 반영하여 여러 하위점수를 한 차원으로 합성함에 따라 평균은 일관되게 유지되는 반면 표준편차가 상대적으로 작아진다. 또한 〈표 9〉의 상관계수 비교에서 다차원 문항반응모형의 총점은 일차원 문항반응모형의 결과 및 서로간에 매우 높은 단순상관(대각선 아래)과 서열상관(대각선 위)을 나타낸다. 이는 모의 자료 분석의 복합구조에서 나타난 기준선

5) 평균 가중치로 계산된 각도이기에 두 값의 합이 정확히 90°를 이루지 못한다.

합성방법과 검사정보함수 방법간의 높은 상관이 실제 자료 분석에서도 동일하게 확인된 것이라 할 수 있다. 또한 앞서 설명된 바와 같이 어휘/문법 검사의 두 하위영역 원점수 상관, 하위점수와 총점의 상관이 모두 상대적으로 높은 특성이 기준선 합성방법과 검사정보함수 방법 총점의 유사성으로 이어진다.

다차원 문항반응모형의 두 가지 총점의 높은 상관에도 불구하고, 보다 세밀한 비교를 위하여 피험자를 10개 집단으로 구분하여 하위점수와 총점과의 관계를 〔그림 2〕에 제시했다. 〔그림 2〕에서 실선 위의 점은 각 방법의 총점 서열에 따라 200명씩 묶인 10개 집단에 대한 하위점수의 위치를 나타낸다. 예를 들어, 집단 5에 대한 기준선 합성방법의 총점 평균은 -0.11이며, 이에 대응하는 하위점수(어휘능력, 문법능력)의 평균은 (-0.11, -0.88)이다. 유사하게 집단 5의 검사정보함수 방법의 총점 평균은 -0.06이며, 이에 대응하는 어휘능력과 문법능력은 (-0.12, -0.06)이다.



(가) 기준선 합성방법 (나) 검사정보함수 방법

〔그림 2〕 10개 피험자 집단의 점수 분포

〔그림 2〕에 나타난 바와 같이 다차원 문항반응모형의 두 가지 총점은 하위점수에 동일한 가중치를 부여하는 45°선(점선)에서 일정정도 이탈하여 위치한다. 그러나 기준선 합성방법이 상대적으로 45°선에 가까우며, 직선의 관계를 유지한다. 이는 기준선 합성방법이 모든 피험자에게 동일한 가중치를 설정하기 때문에 나타난 결과라 할 수 있다. 이에 반하여 검사정보함수 방법에서 총점에 따른 가중치의 변화가 명확히 나타난다. 즉, 오른쪽 그림의 집단 1~집단 3 사이에서 주로 첫 번째 차원의 점수에서 차이가 발생한다면, 집단 3~집단 9 사이에서는 두 번째 차원의 점수의 변화가 총점에 크게 작용한다.

결국 기준선 합성 총점과 검사정보함수 총점의 상관이 매우 높음에도 불구하고(〈표 9〉 참조), 총점에 미치는 하위점수의 상대적 영향력(혹은 중요도)이 두 가지 방법에서 차별적으로 나타나

며, 특히 검사정보함수 방법은 피험자 수준에 따른 하위영역의 상대적 중요도에 대한 보다 세부적 정보를 제공한다.

Ⅴ. 결론 및 논의

다차원 문항반응이론의 검사 차원성와 관련하여, ㄱ) 검사가 다차원적일지라도 피험자가 한 차원에서만 변산을 갖는 경우, ㄴ) 피험자가 둘이상의 차원에서 변산을 가질지라도 검사가 한 차원만을 측정하는 경우, ㄷ) 피험자가 둘 이상의 차원에서 변산을 갖고 검사가 둘 이상의 차원을 측정할지라도, 검사에 포함된 문항벡터가 모두 동일한 방향일 경우, 일차원 문항반응모형과 다차원 문항반응모형은 본질적으로 동일하다(Reckase, 1990, 2009). 이상의 세 가지 조건은 검사의 차원구조와 관련된 것으로, 다차원 문항반응모형이 다양한 능력차원을 측정할 때, 이들을 종합한 총합 점수는 검사의 차원구조에 따라 단일 능력차원을 가정하는 일차원 문항반응모형의 결과와 비교되거나 다양한 방식으로 하위영역의 가중치가 산출된다.

이 연구는 단순구조와 복합구조의 검사 조건에서 모의 자료와 실제 검사자료를 활용하여 다차원 문항반응모형의 두 가지 총점 산출방법(기준선 합성방법과 검사정보함수 방법)을 비교하였다. 연구결과로써, 모의 자료 분석에서 검사의 차원구조가 단순구조에 가까울수록 기준선 합성방법과 검사정보함수 방법의 총점 간 차이가 크게 나타났다. 즉, 개별 문항이 서로 구분되는 능력 차원을 순수하게 측정하는 경향이 강할수록 피험자의 위치에 따라 오차분산을 최소화하는 총점의 방향이 달라지며, 이는 단순구조에서 총점-하위점수 사이에 단일한 선형관계를 가정할 수 없음을 의미한다. 두 번째, 복합구조를 보이는 실제 자료 분석에서 두 가지 방법이 유사한 총점 결과를 제공함(높은 상관, 〈표 9〉 참조)에도 불구하고, 검사정보함수 방법의 총점은 피험자 능력수준에 따라 상대적으로 중요한 역할을 하는 잠재적 차원에 대한 보다 세밀한 정보를 제공하는 특성을 보인다(〔그림 2〕 참조). 즉, 상관이라는 전체 피험자 대상의 통계치에서는 두 방법이 매우 유사하지만, 개별 피험자의 총점에 반영되는 하위점수의 위치는 차별적으로 나타난다.

기준선 합성방법과 검사정보함수 방법의 수리적 차별성은 하위점수와 총점간에 단일한 선형관계를 가정할 수 있는가의 여부이다(Yao, 2010, 2011). 먼저 모든 피험자에게 동일한 하위점수 가중치를 설정하는 기준선 합성방법은 여러 변수의 공통분산을 추출하는 요인분석 방법과 유사하게 이해될 수 있다. 이는 다양한 차원의 정보를 한 개 차원(혹은 적은 수의 차원)으로 투영(projection)함으로써 정보를 요약하는 방법이며, 이를 달리 표현하면, 기준선 합성방법은 검사가 다차원 구조를 가짐에도 불구하고 단일 차원성을 가정하는 일차원 문항반응모형과 논리적으로 유사한 결과를 산출한다. 이에 반하여 검사정보함수 방법은 피험자와 문항간의 상호작용이

피험자 수준에 따라 달라지는 복잡한 상황을 수리 모형에 반영하여 검사정보를 최대화하는 총점을 산출하는 방법이라 할 수 있다. 그러나 검사정보를 최대화하는 총점 산출방법은 신뢰도라는 측면에서 강점을 보임에도 불구하고 피험자에 따라 하위점수의 가중치가 달라진다는 특성을 보인다. 즉, 검사정보함수 방법은 하위점수의 평균 혹은 일관된 하위점수 가중치를 설정하는 전통적 방법과 비교하여 피험자의 수준별 특성을 반영하여 총점을 산출한다는 새로운 관점을 제공하는 혁신적 점수 산출방법(innovative scoring methods)이라 할 수 있다(Yao, 2010).

그러나 모의 자료 분석에서 나타나 바와 같이 하위점수의 독자성이 강하게 나타나는 단순구조에서 최대 검사정보를 이용한 총점 산출은 상대적으로 주의가 요구된다. 복합구조와 비교하여 단순구조에서는 능력수준에 따라 특정 능력차원(하위점수)이 총점에 미치는 영향이 과대하게 설정되며, 이는 피험자에 따른 검사 총점의 형평성에 대한 현실적 문제로 이어진다. 이러한 차원구조에 따른 차이는 다차원 문항반응모형의 하위점수 신뢰도가 단순구조에서 높아진다는 것과 연관된 것으로(민경석, 2011), 검사의 다차원성을 반영한 측정모형을 통한 점수 산출에서 차원구조가 중요한 검사 특성으로 고려되어야할 것이다.

다차원 문항반응이론에 대한 학술적 연구와 실제적 적용에 대한 논의가 확대됨에도 불구하고 검사자료의 차원 수, 차원구조, 하위영역 간의 상관 등에 대한 명확한 통계적 방법이 제시되지 못하고 있다(Reckase, 2009). 이 연구는 두 가지 이론적 차원구조가 다차원 문항반응모형의 총점 산출에 미치는 영향을 분석하면서 가장 단순한 형태인 2차원성을 가정했으며, 또한 실제 자료의 분석은 복합구조에 제한된다. 검사 점수 산출방법은 측정모형의 현실적 적용을 위한 중요한 요소로, 차원구조와 직접적으로 관련된 검사의 차원 수, 하위점수의 상관 등과 같은 검사 조건과 다양한 측정모형(예, 위계적 문항반응모형, 2요인 모형 등)을 고려하여 하위점수–총점 관계에 대한 지속적인 연구가 필요할 것이다.

참 고 문 헌

강태훈(2010). **다차원문항반응이론에 기반한 종합 및 하위능력 모수 추정방법간 비교 연구**. 한국교육학회 춘계학술대회(한국교육평가학회). 한국교원대학교.

김성숙, 상경아, 이상아(2009). 국가수준 학업성취도 평가 개선 연구. **한국교육과정평가원 연구 리포트**.

김신영(1999). 대학수학능력시험의 영역별 변환표준점수 반영비율에 관한 연구. **교육과정평가연구, 2**(1), 289-299.

민경석(2011). 문항반응이론에 기반한 하위점수 산출방법 비교. **교육평가연구, 24**(3), 799-817.

Ackerman, T. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18,* 257-276.

ACT (2007). *ACT Technical Manual*. Iowa City, IA: ACT.

ACT (2008). *The ACT User Handbook 2008/2009*. Iowa City, IA: ACT.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Boughton, K. A., Yao, L, & Lewis, D. M. (2006). *Reporting diagnostic subscale scores for tests composed of complex structure.* Paper presented at the 2006 National Council on Measurement in Education. San Francisco, California.

DeMars, C. E. (2005). *Scoring subscales using multidimensional item response theory models*. Poster presented at the annual meeting of the American Psychology Association. Available at http://www. *jmu.edu/assessment/wm_library/subscaledemo.doc*.

Oosterhof, A. C. (1987). Obtaining intended weights when combining students' scores. NCME Instructional Module.

Reckase, M. D. (1990). Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, NJ.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York. NY: Springer.

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.

Sinharary, S., Haberman, S., & Puhan, G. (2008). Subscores based on classical test theory: To report or not to report. *Educational Measurement, 26*(4), 21-28.

Wang, W.-C., Chen, P.-H., & Cheng, Y-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*(1), 116-136.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47*(3), 339-360.

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement, 35*(1), 48-66.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83-105.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement, 30,* 469-492.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG.* [Computer software and manual]. Lincolnwood. IL: Scientific Software International.

· *논문접수 : 2012-09-01/ 수정본접수 : 2012-10-11/ 게재승인 : 2012-10-17*

ABSTRACT

Comparison of Reference Composite Method and Maximizing Test Information Function Method to Calibrate Overall Scores in MIRT

Kyung-Seok Min
(Associate Professor, Sejong University)

Multidimensional IRT models have been popular as the tests tend to measure multiple constructs rather than one latent variable. While there were various studies to produce more reliable sub-scores in multidimensional IRT models, the research on overall scores was somewhat limited. The purpose of this paper was to compare two MIRT calibration methods of overall score such as reference composites and maximizing test information function method. For the comparison of two methods, simulated data with approximate simple structure and complex structure, and real data were analyzed. As a result, a simple linear relationship between sub-scores and overall scores could not be assumed expecially in the simple structure. Also, the test information function method provided more detailed examinees' ability information in the complex structure compared with the reference composit method. In addition, limitations of the study and directions for further research were discussed.

Key Words : multidimensional IRT models, reference composite, test information function, simple structure, complex structure