

국가수준 학업성취도 평가 결과 분석에 적합한 인지진단모형 탐색: Fusion 모형과 DINA 모형의 비교¹⁾

김 희 경(한국교육과정평가원, 연구위원)*
박 찬 호(한국교육과정평가원, 부연구위원)

《 요 약 》

본 연구에서는 인지진단모형을 활용하여 국가수준 학업성취도 평가 결과를 분석하고, 학생들에게 프로파일 형태의 상세한 피드백을 줄 수 있는 방안을 마련하고자 자료 특성에 가장 적합한 모형을 탐색하였다. 선행연구를 통해 대규모 평가 분석에 자주 활용되는 것으로 파악된 인지진단모형인 Fusion Model과 DINA Model 중에서 학업성취도 평가 자료에 더 적합한 모형을 탐색하고자 하였다. 연구 결과, 약 60만명의 대단위 학생 자료를 포함하고 있는 학업성취도 평가 자료에 인지진단모형 분석을 위한 프로그램을 적용하여 개별 학생의 세부 인지요소 숙달 수준에 대한 프로파일 정보를 산출할 수 있는 가능성을 확인하였다. 또한, Fusion Model과 DINA Model을 비교한 결과 인지요소 숙달 수준을 숙달/미숙달의 2단계가 아닌 상/중/하의 3단계로 분류하는 경우, DINA Model은 '상'이 0% 또는 '하'가 0%에 가깝게 나타나는 등 인지요소 숙달 수준이 편중되는 경향을 보였다. 따라서, 국가수준 학업성취도 평가 자료를 인지진단모형을 통해 2단계 이상의 숙달수준으로 분류하는 경우 Fusion Model이 더 적합한 것으로 나타났다.

주제어 : 국가수준 학업성취도 평가, 인지진단모형, Fusion Model, DINA Model, Q행렬

1) 이 논문은 한국교육과정평가원에서 수행한 '인지진단모형을 적용한 학업성취 프로파일 분석 및 결과 보고 방안'(김희경, 한정아, 최숙기, 김부미, 2012) 연구 결과의 일부를 발췌하여 수정·보완한 것임.

* 제1저자 및 교신저자, heekyoung@kice.re.kr

I. 서론

인지진단이론은 검사에 의해 측정되는 여러 가지 지식 및 기능에 대해 학생이 습득하고 있는 지 여부를 상세하게 파악하여 학생의 학력을 총체적인 관점에서가 아니라, 다양한 측면에서 파악하여 학습에 실질적인 도움을 주고자 개발된 이론이다(Tatsuoka, 1985; DiBello et al., 1995; Hartz, 2002). 인지진단이론은 학생들의 평가 결과를 교수·학습 과정 개선에 직접적으로 연계할 수 있는 방안을 마련하는 데 유용하며, 최근에는 이러한 필요성과 맞물려 다양한 인지진단모형들이 개발되었다(Rupp & Templin, 2008). 현재 개발된 인지진단모형은 Rule Space Model, Fusion Model, DINA Model, NIDA Model, Generalized Diagnostic Model 등 다양하며, Fu & Li(2007)는 62종의 인지진단모형에 대해 리뷰하기도 하였다. Tatsuoka(1983)에 의해 개발된 Rule Space Model은 Q행렬의 도입 등 다른 인지진단모형들이 개발되는 데 영향을 준 선도적인 모형이라 할 수 있다.

근간에는 평가 결과에 기반한 학습 개선 방향 탐색이라는 실제적인 요구에 부합하기 위하여 인지진단모형을 활용한 국가수준, 혹은 대규모 평가 결과 분석에 관한 실천적 사례들이 등장하기 시작하였다. 실제로 해외의 대규모 평가에서 인지진단모형을 활용해 결과 분석을 시도한 대표적 사례로 미국의 PSAT/NMSQT(Preliminary SAT/National Merit Scholarship Qualifying Test)²⁾의 'Score Report Plus'TM에서 제공하는 성적표를 들 수 있다(College Board, 2012).

특히 대단위 자료 분석에 널리 사용되고 있는 인지진단모형으로 Fusion Model과 DINA Model을 들 수 있다. 다시 말해 Fusion Model과 DINA Model은 현재 가장 널리 쓰이는 대표적인 인지진단모형이다. Fusion Model과 DINA Model은 공통적으로 문항 풀이를 위해 요구되는 인지요소들을 모두 숙달해야 정답을 맞힐 수 있다는 결합적 가정(conjunctive assumption)에 기초한다. 또한 두 모형은 한 인지요소에 대한 습득이 지나치게 훌륭하다고 할지라도 다른 인지요소를 습득하지 못해 발생하는 학업적 결핍을 보상하지 못한다는 전제에 기초하므로, 필요한 인지요소를 하나라도 습득하지 못하면 문항을 틀리게 된다는 비보상적 모형(non-compensatory model)에 속한다.

이렇게 결합적 가정에 기초한 비보상적 모형의 장점은 수학적으로 비교적 간단하다는 점이다. 수학적으로 너무 복잡한 인지진단모형의 경우, 자료의 크기가 크면, 실제로 추정해야 하는 모수의 수가 너무 많아 대단위 자료에는 적용하기 힘든 제한점이 있는 반면 Fusion Model과 DINA Model은 수학적으로 덜 복잡하여 많은 수의 학생들의 숙달/미숙달 정보를 분석하는 데 있어 무리가 없고, 이해와 해석이 유용하다는 장점이 있다.

2) SAT를 대비하기 위한 모의고사로서 SAT와 같은 유형의 문항들로 구성되어 있다. 매년 10월에 시행되며, 우리나라의 모의 수능과 같은 성격의 시험이라고 할 수 있다.

선행 연구들에서는 대부분 표집자료를 대상으로 연구를 수행하였고, 전수자료를 대상으로 인지진단모형을 적용한 사례는 국내외를 통틀어 찾기 힘들다. 국내에서도 대부분 인지진단모형을 적용한 분석을 학급 단위로 시행하거나 몇 개 학교를 대상으로 소규모 자료를 사용하여 표본 크기를 언급하지 않은 연구가 대부분이었다. 대규모 평가인 국가수준 학업성취도 평가 자료를 사용한 김성훈, 송미영(2011)과 송미영 외(2011)의 연구에서도 표본 크기는 각각 5,185명, 10,265명이었다.

본 연구에서는 국가수준 학업성취도 평가 자료를 대상으로 평가 결과를 분석하고 보고하는데 적합한 인지진단모형을 탐색하고자 한다. 첫째, 선행연구 검토에 의하면, Fusion Model과 DINA Model이 대규모 자료 분석에 자주 이용된 모형임을 파악하였고, 따라서 두 가지 모형을 중심으로 국가수준 학업성취도 평가 자료에 더 선호되는 인지진단모형을 탐색하고자 한다. 둘째, 지금까지 전수자료를 대상으로 인지진단모형을 적용한 사례가 국내외를 통틀어 없기 때문에 자료의 크기가 60만명 이상인 대규모 자료 분석에 인지진단모형을 적용한 분석이 가능한지 파악하고자 한다.

II. Fusion Model과 DINA Model의 특징

현재까지 개발된 다양한 인지진단모형 가운데 대규모 자료 분석에 널리 사용되고 있는 인지진단모형은 Fusion Model과 DINA Model이라고 할 수 있다. <표 1>는 선행연구 검토를 토대로 Fusion Model과 DINA Model의 특징을 비교하여 제시하고 있다.

<표 1> Fusion Model과 DINA Model의 비교

모형	추정모수	서답형 문항 처리	프로그램	대단위 자료 적용 예시
Fusion	<ul style="list-style-type: none"> • 문항모수 : π_i^* (난이도), r_{ik}^* (변별도), c_i (Q행렬 의존도) • 학생모수 : α_j (인지요소 숙달 확률), η_j (잔여 능력) 	서답형 문항의 다분점수를 그대로 처리 가능	Arpeggio	<ul style="list-style-type: none"> • Hartz(2002) : PSAT/NMSQT, ACT • Henson & Templin(2004) : NAEP • Lee & Sawaki(2009) : TOEFL iBT • Jang(2009): TOEFL • Li(2011) : MELAB(Michigan English Language Assessment Battery)

모형	추정모수	서답형 문항 처리	프로그램	대단위 자료 적용 예시
DINA	<ul style="list-style-type: none"> • 문항모수 : s_i (이탈), g_i (추측) • 학생모수 : α_j (인지요소 숙달 확률), η_j (문항 숙달여부) 	다분점수를 이분점수 (0/1)로 리코딩하여 처리	OX	<ul style="list-style-type: none"> • de la Torre(2008): NAEP • Huebner 외(2009): GMAT • Chiu 외(2011): PIRLS

1. Fusion Model

초기의 Fusion Model은 Hartz(2002)가 개발한 모형으로 학생 j 가 문항 i 의 정답을 맞힐 확률을 식 (1)과 같이 표현한다.

$$P(X_{ij} = 1 \mid \alpha_j, \eta_j) = \pi_i^* \prod_{k=1}^k (r_{ik}^{*(1-\alpha_{jk}) \times q_{ik}}) P_{c_i}(\eta_j) \quad (1)$$

최근 Fusion Model은 서답형 문항 분석이 가능하도록 확장되었다. 부분점수 득점이 가능한 서답형 문항이 포함된 경우에는 좀 더 복잡한 수리적 모형이 필요하다. 예를 들어 배점이 4점인 서답형 문항에서는 0, 1, 2, 3, 4의 5개의 다분점수가 가능할 것이다. Fu(2005)는 Fusion Model을 다분문항에도 적용할 수 있도록 확장하였는데, 확장된 모형에서는 M 개의 다분점 점수(0, 1, ..., M_i-1)가 존재하는 다분문항의 경우, 학생 j 가 문항 i 에서 m 번째 점수 이상을 득점할 확률을 식 (2)와 같이 표현하였다.

$$P_{ijm}^* = P(x_{ij} \geq m \mid \alpha_j, \eta_j) = \begin{cases} 1 & m=0 \\ \pi_{im}^* \prod_{k=1}^K r_{imk}^{*(1-\alpha_{jk}) \times q_{ik}} P_{cim}(\eta_j) & m=(1, \dots, M_i-1), \end{cases} \quad (2)$$

결국, 정확히 m 번째 점수를 득점할 확률은 식 (3)과 같이 표현할 수 있다.

$$P_{ijm} = P(x_{ij} = m \mid \alpha_j, \eta_j) = \begin{cases} P_{ijm}^* - P_{ij(m+1)}^* & m=(0, 1, \dots, M_i-2) \\ P_{ijm}^* & m=M_i-1 \end{cases}, \quad (3)$$

π_{im}^* 은 문항 i 의 풀이에 요구되는 모든 인지요소를 숙달한 학생이 자신이 숙달한 인지요소를 정확히 사용하여 문항 i 에서 적어도 m 번째 점수를 득점할 확률을 말한다. r_{imk}^* 은 $\frac{P(Y_{ijk} \geq 1 \mid \alpha_{jk}=0)}{P(Y_{ijk} \geq 1 \mid \alpha_{jk}=1)}$ 로 표현되며 인지요소 k 를 숙달하지 못했을 때 문항 i 에서 적어도 m 번째

점수를 득점할 확률에 대한 인지요소 k 를 숙달했을 때 문항 i 에 적어도 m 번째 점수를 득점할 확률의 비율이다. 또한 q_{ik} 는 Q행렬에 규정되어 있는 문항 i 의 풀이에 요구되는 인지요소 k 를 말하고, c_i 는 문항이 Q행렬에 명시되지 않은 능력에 대해 어느 정도 의존하고 있는지를 나타내며, α_j 는 학생모수로서 학생 j 가 각 인지요소를 숙달했는지를 나타내는 벡터로 인지요소 k 를 숙달하면 1, 그렇지 않으면 0의 값을 가진다.

2. DINA Model

DINA(The Deterministic Inputs, Noisy “And” gate) Model에서는 학생의 문항에 대한 반응은 문항을 맞히기 위해 요구되는 인지요소들을 학생이 습득하고 있는지에 따라 결정된다고 가정한다(Haertel, 1990). 학생의 문항에 대한 실제 반응은 기대 반응과 다를 수 있는데, 학생의 기대 반응이 정답이었는데 실제 반응이 오답으로 나타나면 슬립(slip)한 것으로 표현하고, 기대 반응이 오답이었는데 실제 반응이 정답으로 나타나면 추측(guess)한 것으로 개념화한다. DINA Model에서는 슬립과 추측을 실제와 기대 사이에 혼란을 일으키는 잡음(noise)이라 규정하고 있다.

DINA Model에서는 학생 j 가 문항 i 의 정답을 맞힐 확률을 식 (4)와 같이 표현한다.

$$P(X_{ij} = 1 \mid \alpha_j) = (1 - s_i)^{\eta_{ij}} g_i^{1 - \eta_{ij}} \quad (4)$$

위 식에서 $\eta_{ij} = \prod_{k=1}^K \alpha_{ik}^{q_{ik}}$ 로서 학생 j 가 문항 i 의 정답을 맞히기 위한 모든 인지요소를 숙달했는지를 나타낸다(0=미숙달, 1=숙달). 즉, 학생 j 가 k 개의 인지요소 중에서 문항 i 에 필요한 모든 인지요소를 숙달했으면 ‘1’, 하나라도 숙달하지 못했으면 ‘0’의 값이 주어진다. α_j 는 학생모수로서 학생 j 가 각 인지요소를 숙달했는지를 나타내는 벡터이다. $P(X_{ij}=1|\underline{\alpha}, s, g)$ 는 각 인지요소에 대한 숙달 여부 벡터 $\underline{\alpha}$ 를 지닌 학생 j 가 슬립모수 s , 추측모수 g 를 가진 문항 i 의 정답을 맞힐 확률을 나타낸다. 슬립모수 s_i 는 학생이 문항 i 의 풀이에 요구되는 모든 인지요소를 숙달하였음에도 불구하고 그 문항을 틀릴 확률이다. 반면에 추측모수 g_i 는 학생이 문항 i 의 풀이에 요구되는 모든 인지요소를 숙달하지 못했음에도 불구하고 그 문항을 맞힐 확률이다. Q행렬이 잘못 정의된 경우, 또는 학생들이 그 문항에서 기대하고 있는 문제풀이 전략과 다른 전략을 시도한다거나, 모형 적합도가 떨어지는 경우 s_i 와 g_i 의 값이 높아질 수 있다(Junker & Sijtsma, 2001). 학생의 잠재 또는 기대 반응 패턴 η_{ij} 를 알고 문항모수인 슬립모수(s_i)와 추측모수(g_i)를 알면, 문항의 정답 확률이 추정된다.

Ⅲ. 연구 방법

1. Q행렬

인지진단모형을 적용하기 위해서는 검사를 구성하는 문항들과 그 문항의 정답을 맞추기 위해 요구되는 인지요소들 간의 관계를 도출해야한다. 인지진단모형을 적용하기 위해 문항과 인지요소 간의 관계를 명시해놓은 $k \times n$ (k =인지요소의 수, n =문항의 수)의 행렬을 Q행렬이라고 한다.

본 연구에서는 2011년 학업성취도 평가 국어과, 수학과와 Q행렬 개발을 위해 국어과 내용 전문가 8인, 수학과 내용 전문가 6인이 Q행렬 개발 과정에 참여하였다. 먼저 2011년 국가수준 학업성취도 평가들에서 제공하는 성취기준을 토대로 실제로 문항을 풀어보면서 초6, 중3, 고2의 국어, 수학과 학업성취도 평가에서 측정하고 있는 공통의 인지요소를 추출하였고, 다음에 인지요소와 관련된 문항을 연결하였다. Q행렬 초안이 개발된 후에는 통계적 타당화 검증 결과(자카드 계수 및 다중회귀분석)³⁾를 기반으로 교과 전문가들의 판단 아래 수정·보완을 반복하였다.

2. 분석 자료 및 방법

2011년 학업성취도 평가 자료에서 분석에 포함된 사례 수는 <표 2>에 제시하였다. 고등학교 2학년은 전문계를 제외한 일반계 학생만을 포함하여 분석하였다.

본 연구에서 Fusion Model을 적용하기 위해 Arpeggio version 3.1(Dibello & Stout, 2010)을 사용하였는데, 이 프로그램은 서답형 문항의 부분점수를 처리할 수 있도록 개발된 것이다. DINA Model은 현재까지 개발된 프로그램이 이분적(0/1)으로 코딩된 자료에만 적용 가능하기 때문에 서답형 문항의 부분점수를 0/1로 변환하여 리코딩한 자료를 사용하여 분석하였다. de la Torre(2009)는 OX(Doornik, 2003) 환경에서 실행 가능한 DINA Model 분석용 코드를 작성하였으나 de la Torre의 프로그램은 본 연구에서 사용한 대규모 자료를 처리할 수 없었다. 따라서 본 연구에서는 WinBUGS(Spiegelhalter, Thomas, Best, & Lunn, 2003)를 이용하여 각 자료의 표본으로부터 얻은 문항 모수를 고정 한 채 피험자의 인지요소 숙달확률을 계산하였다.

3) 자카드 계수(Jaccard index)는 Park(2008)의 연구에서 Q행렬 타당화를 위해 사용되었는데, Q행렬에 나타난 인지요소 간의 유사성을 산출하는데 유용하다. 자카드 계수는 Jaccard similarity coefficient라고도 하며, $J(A, B) = |A \cap B| / |A \cup B|$ 로 표현된다(Hannig, 2004 재인용). 다중회귀분석은 김성훈(2005)과 Buck & Tatsuoka(1998)이 Q행렬 타당화를 위한 통계적 방법으로 사용하였다. 문항의 정답률을 종속변인, Q행렬을 독립변인으로 다중회귀분석을 수행하여 인지요소의 적절성 여부를 검토하는 데 참고자료로 쓰일 수 있다.

〈표 2〉 인지진단모형 분석 대상

교과	사례 수		
	초등학교 6학년	중학교 3학년	고등학교 2학년
국어	578,020	628,200	486,675
수학	576,166	628,369	469,797

IV. 연구 결과

1. Q행렬 최종안

본 연구에서 2011년 학업성취도 평가를 대상으로 초·중·고 국어과, 수학과 Q행렬 최종안을 개발하였다. 지면상 여기에서는 고2 국어과와 초6 수학과 Q행렬을 〈표 3〉과 〈표 4〉에 각각 제시하였다.

〈표 3〉 2011년 학업성취도 평가 고2 국어 Q행렬 최종안

인지요소 문항	국어 지식	단어 의미 파악	문장 관계 파악	내용 파악		추론 및 탐구					비판		표현하기	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	0	0	1	1	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	1	0	1	0
6	0	0	0	0	0	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	0	1	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0
11	1	0	0	0	0	0	0	0	0	1	0	0	0	0
12	0	0	0	0	1	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	1	0	0	0	0	0	0	0	0

인지요소 문항	국어 지식	단어 의미 파악	문장 관계 파악	내용 파악		추론 및 탐구					비판		표현하기	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
14	0	0	0	0	0	0	0	0	0	0	0	1	0	0
15	0	0	0	1	0	0	0	0	0	0	1	0	0	1
16	0	1	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	1	0	0	0	0	0
18	0	0	0	0	1	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	1	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	1	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	1	0	0
22	0	0	0	1	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	1	0	0	0	0	0	0	0
24	0	0	1	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	1	0	0	0	1	0	0
26	0	0	0	0	0	0	1	0	0	0	0	0	0	0
27	0	0	0	0	0	0	1	0	0	0	0	0	0	0
28	0	0	0	0	0	0	1	0	0	0	0	0	0	0
29	0	0	0	0	0	0	1	0	0	0	0	0	0	0
30	0	0	0	0	0	0	1	0	0	0	0	0	0	0
31(서1)	1	0	0	0	0	0	0	1	0	0	0	0	0	0
32(서2)	1	0	0	0	0	0	0	0	0	0	0	0	0	1
33(서3)	1	1	1	0	1	0	0	0	0	0	0	0	0	1
34(서4)	0	1	0	0	1	0	0	1	0	0	0	0	0	0
35(서5)	0	0	0	0	0	1	1	0	0	0	0	0	0	0
36(서6)	0	1	1	0	0	0	0	0	0	0	0	0	0	1

〈표 4〉 2011년 학업성취도 평가 초6 수학 Q행렬 최종안

인지요소 문항	계산하기		이해하기		추론하기	표현하기			해석하기			
	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	0	1	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	1	0	0	0	0	0	0	0	1

문항 인지요소	계산하기		이해하기		추론하기	표현하기			해석하기			
	1	2	3	4	5	6	7	8	9	10	11	12
5	1	0	0	1	0	0	0	0	0	0	0	0
6	1	0	0	1	0	0	0	0	0	0	1	0
7	0	1	0	0	0	0	0	0	0	0	1	0
8	0	0	0	0	0	0	0	0	0	0	1	0
9	0	0	0	1	0	0	1	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0	1	0
11	0	0	0	1	0	0	0	0	0	0	0	0
12	0	0	1	0	0	0	0	0	0	0	0	0
13	1	0	0	0	0	0	0	0	1	0	0	0
14	0	0	0	1	0	0	0	0	0	0	0	0
15	1	0	0	1	0	0	1	0	1	0	0	0
16	0	0	0	0	0	0	1	0	0	0	0	0
17	0	0	0	1	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	1	0	0
19	0	0	0	0	0	1	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	1	0
21	1	0	1	0	0	0	0	0	1	0	0	0
22	0	0	0	1	0	0	0	0	0	0	0	1
23	0	0	0	0	0	0	0	0	1	0	0	0
24	1	0	0	0	0	0	0	0	0	0	0	0
25	1	0	0	1	0	0	0	0	1	0	0	0
26(서1)	0	0	0	1	0	0	0	0	0	0	0	0
27(서2)	1	0	0	1	0	0	0	1	0	0	0	0
28(서3)	0	0	0	1	0	0	0	1	0	0	0	0
29(서4)	0	0	0	0	1	1	0	1	1	0	0	0

2. Fusion Model과 DINA Model의 비교

Fusion Model과 DINA Model을 학업성취도 평가 자료에 적용하였을 때, 산출되는 학생들의 학업성취 프로파일 결과를 비교하였다. 초6, 중3, 고2 국어과, 수학과 자료를 대상으로 Fusion Model과 DINA Model의 결과를 비교하였으나 지면상 고2 국어과와 초6 수학과와 결과를 제시하기로 한다.

인지진단모형에서 추정되는 학생모수인 α_j (인지요소 숙달여부)는 각 학생이 해당 인지요소를 숙달하고 있을 확률(Posterior Probability of Mastery: PPM)이며 0~1 범위 내의 값을 갖는다. Lee & Sawaki(2009), Li(2011), 김수진 외(2008), 송미영 외(2011)는 연속변수로 계산되는 PPM을 0.5를 기준으로 숙달/미숙달을 분류하였다.

0.5를 기준으로 PPM이 0.5 이상은 숙달, PPM이 0.5 미만은 미숙달로 구분하였을 때, 인지요소별 숙달/미숙달 비율을 두 가지 모형에서(Fusion vs. DINA) 비교하였다.

〈표 5〉 고2 국어 인지요소별 숙달/미숙달 비율(분류 기준=0.5)

인지요소	Fusion		DINA	
	숙달	미숙달	숙달	미숙달
1	87.72	12.28	85.94	14.06
2	87.55	12.45	87.53	12.47
3	90.73	9.27	84.11	15.89
4	91.10	8.90	71.19	28.81
5	89.34	10.66	75.54	24.46
6	88.47	11.53	73.04	26.96
7	93.48	6.52	63.51	36.49
8	86.15	13.85	76.84	23.16
9	92.93	7.07	70.93	29.07
10	93.51	6.49	66.91	33.09
11	93.44	6.56	55.56	44.44
12	94.14	5.86	66.89	33.11
13	94.21	5.79	83.01	16.99
14	91.08	8.92	87.42	12.58

고2 국어과 14개 인지요소에 대한 인지요소별 숙달/미숙달 비율을 두 가지 모형을 적용하여 분석한 결과를 〈표 5〉에 제시하였다. 〈표 5〉에서 DINA Model을 통해 산출된 숙달의 비율은 Fusion Model을 통해 산출된 경우보다 낮고 미숙달의 비율은 높게 나타났다. 특히, 두 가지 모형을 사용하였을 때, 미숙달 비율 간 차이가 크게 나타난 인지요소는 인지요소 7(30%p 차이), 인지요소 11(38%p 차이), 인지요소 12(27%p 차이)였는데, 이들 인지요소의 특징은 인지요소 7과 12는 각각 서답형 문항이 2개씩 포함된 문항으로 서답형 문항과 관련된 인지요소였다.

이러한 결과는 DINA Model을 적용하기 위해 서답형의 부분점수를 반영하지 못하고, 0/1의 이분점수를 사용한 점에 영향을 받은 것으로 해석된다. 즉, DINA Model 적용 시 학생들의 서답형 문항에서의 득점을 0/1로 간소화하면 학생들의 서답형 문항에서의 부분점수

특점을 제대로 반영하지 못해 그들의 능력을 과소 추정하는 원인이 된다고 볼 수 있다.

PPM을 근거로 인지요소 숙달수준에 대한 의사결정을 하는 방법은 0.5를 기준으로 양분하는 방법이외에 다른 방법을 사용할 수 있다. Jang(2005)은 (0.4, 0.6)을 기준으로 PPM이 '0.4 미만'인 경우에 미숙달, '0.4 이상~0.6 미만'인 경우에 불확실, '0.6 이상'인 경우 숙달로 분류하기도 하였다. 이는 숙달수준을 양분하기보다 2수준 이상으로 분류한 경우인데, Roussos et al.(2007)은 PPM을 분류할 때, 숙달 또는 미숙달과 같이 극단적인 2가지 수준으로 확정하는 것보다 2수준 이상으로 분류할 것을 제안하였다. 또한 (0.2, 0.8)을 기준으로 분류하는 경우, 분류일치도의 일종인 Cohen's kappa 지수를 0.6 이상⁴⁾으로 유지할 수 있다고 하였다.

국가수준 학업성취도 평가에서는 교육과정 목표 도달정도에 따라 학생들의 해당 교과에서의 성취수준을 '우수학력', '보통학력', '기초학력', '기초학력 미달'의 4단계 성취수준으로 구분하는데, 목표도달도가 '0.8 이상'인 경우 우수학력, '0.2 미만'인 경우를 기초미달로 분류하므로 본 연구에서도 인지요소별로 (0.2, 0.8)을 기준으로 '상'과 '하'로 분류해보았다.

〈표 6〉 고2 국어 인지요소별 상/중/하 비율(분류 기준=0.2, 0.8)

인지요소	Fusion			DINA		
	상	중	하	상	중	하
1	78.58	16.55	4.86	81.64	12.31	6.05
2	80.82	11.81	7.37	85.67	4.44	9.89
3	79.10	16.06	4.85	81.15	4.99	13.87
4	80.38	14.74	4.88	60.12	15.94	23.95
5	81.48	13.47	5.05	69.36	12.94	17.69
6	85.56	5.83	8.61	73.04	0.00	26.96
7	77.90	17.73	4.37	51.14	26.56	22.30
8	80.10	15.06	4.84	62.67	27.93	9.40
9	73.41	22.20	4.38	70.93	0.00	29.07
10	80.40	15.25	4.34	66.88	15.55	17.57
11	81.80	13.86	4.33	54.47	19.03	26.50
12	84.10	11.56	4.33	66.89	4.1	29.01
13	83.19	12.49	4.32	0.00	95.48	4.53
14	87.61	6.29	6.11	70.83	28.95	0.22

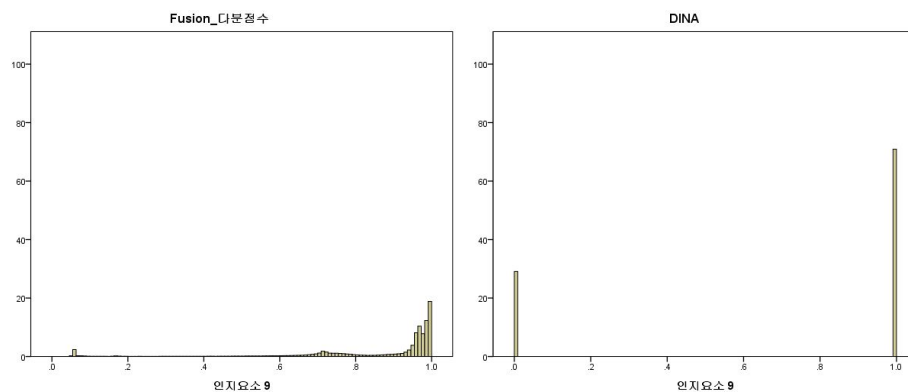
〈표 6〉는 고2 국어 자료를 분석한 결과 산출된 PPM의 수치가 0.2 미만은 '하', 0.2 이상~0.8 미만은 '중', 0.8 이상은 '상'으로 분류한 결과를 Fusion Model과 DINA Model의 경우를

4) Cohen's kappa 지수가 0.6인 경우 '무척 높은'(substantial) 분류일치도를 보인다고 할 수 있다 (Landis & Koch, 1977).

비교한 것이다. Fusion Model을 적용하였을 때, 각 인지요소를 ‘상’의 수준으로 습득하였다고 분류된 비율은 73.4~87.6%로 2011년 고2 국어과 학업성취도 평가에서 ‘우수학력’으로 구분된 학생 비율(33.0%)보다 높았다. 하지만 한 가지 특정 인지요소에서 우수한 것이 전체적으로 우수한 것보다 쉽기 때문에 국어과 전체 우수 비율인 33.0%보다 개별 인지요소에서 우수한 비율이 높은 것은 자연스러운 것으로 보인다. 반면 2011년 학업성취도 평가의 고2 일반계 학생들의 국어과 ‘기초학력 미달’ 비율은 2.0%였는데, Fusion Model에서 산출된 각 인지요소에서의 ‘하’ 비율은 5.83~22.20%로 나타났다.

〈표 6〉에서 음영으로 표시된 부분은 특정 수준에 해당하는 학생의 비율이 0%로 나타난 경우이다. DINA Model을 적용하였을 때, 인지요소 6과 인지요소 9는 ‘중’ 수준의 학생이 0%인 것으로 나타났고, 인지요소 13은 ‘상’ 수준의 학생이 0%로 나타났다.

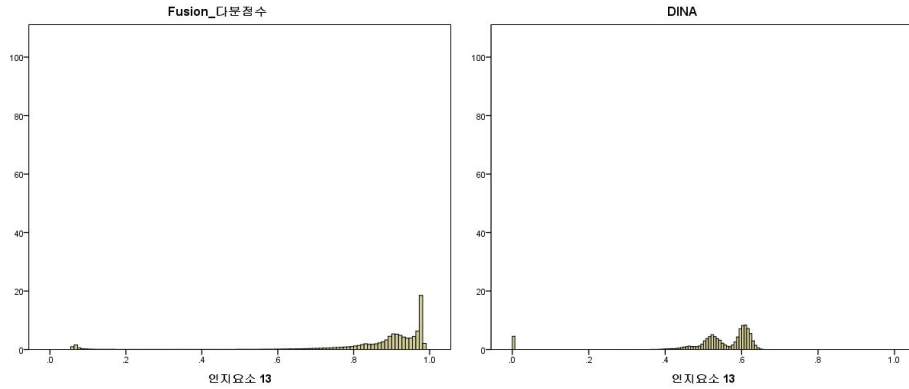
이러한 결과를 토대로 DINA Model의 경우에는 PPM의 분포가 편중되는 경향이 있는 것으로 나타났다. 이는 학업성취도 평가 자료의 인지요소별 프로파일을 숙달/미숙달 이상의 여러 단계로 구분하는 경우에 DINA Model이 한계점이 있음을 시사한다.



〔그림 1〕 고2 국어 인지요소 9의 PPM 분포

인지요소 9에서 DINA Model을 적용한 경우 ‘중’ 수준의 학생이 0%로 산출되었기 때문에 [그림 1]에서는 두 가지 모형을 적용하였을 때의 PPM 분포를 제시하였다. Fusion Model과 달리 DINA Model의 경우 PPM이 0과 1의 극단치에 분포하고 있는 것을 볼 수 있다.

또한, 인지요소 13에서는 DINA Model을 적용한 경우 ‘상’ 수준의 학생이 0%로 산출되었기 때문에 두 가지 모형을 적용하였을 때의 PPM 분포를 [그림 2]에 제시하였다. DINA Model의 경우 PPM이 0과 중간 점수대 부분에만 분포하고 있는 것을 볼 수 있다.



[그림 2] 고2 국어 인지요소 13의 PPM 분포

〈표 7〉 초6 수학 인지요소별 숙달/미숙달 비율(분류 기준=0.5)

인지요소	Fusion		DINA	
	숙달	미숙달	숙달	미숙달
1	47.78	52.22	69.41	30.59
2	74.52	25.48	79.43	20.57
3	73.18	26.82	47.01	52.99
4	76.70	23.30	78.10	21.90
5	81.53	18.47	70.07	29.93
6	84.19	15.81	93.11	6.89
7	53.19	46.81	53.97	46.03
8	77.58	22.42	77.58	22.42
9	82.35	17.65	71.77	28.23
10	83.18	16.82	84.10	15.90
11	73.52	26.48	72.67	27.33
12	83.45	16.55	84.91	15.09

초6 수학과 12개 인지요소에 대한 인지요소별 숙달/미숙달 비율을 두 가지 모형에서 비교하여 〈표 7〉에 제시하였다. 〈표 7〉에서 Fusion Model과 DINA Model의 결과를 비교했을 때, 일관적으로 어느 한쪽의 숙달 비율이 높거나 낮은 패턴이 나타나지 않았다. 인지요소 1, 2, 4, 6, 7, 10, 12에서는 DINA Model의 숙달 비율이 더 높았고, 나머지 인지요소에서는 Fusion Model의 숙달 비율이 더 높았다. 인지요소 8의 경우에는 Fusion Model과 DINA Model의 숙달 비율이 77.58%로 동일하였다. 두 가지 모형 간 미숙달 비율 차이가 가장 큰 경우는 인지요소 1(22%p 차이)과 인지요소 3(26%p 차이)이었으며, 이들 인지요소는 공통적으로 서답형 문항을 2개씩 포함한 문항이었다.

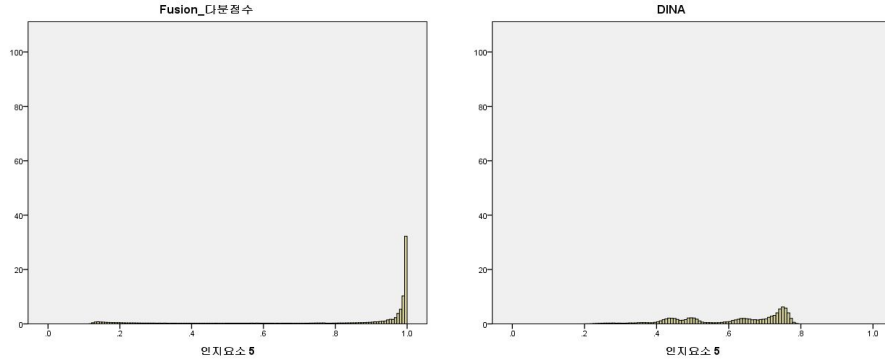
이러한 결과는 국어과의 결과에서와 마찬가지로 DINA Model을 적용하기 위해 서답형의 부분점수를 반영하지 못하고, 0/1의 이분점수를 사용한 점에 영향을 받은 것으로 해석된다. 즉 DINA Model 적용 시 학생들의 서답형 문항에서의 득점을 0/1로 간소화하면 학생들의 서답형 문항에서의 부분점수 득점을 제대로 반영하지 못해 그들의 능력을 과소 추정하는 원인이 된다고 볼 수 있다.

〈표 8〉 초6 수학 인지요소별 상/중/하 비율(분류 기준=0.2, 0.8)

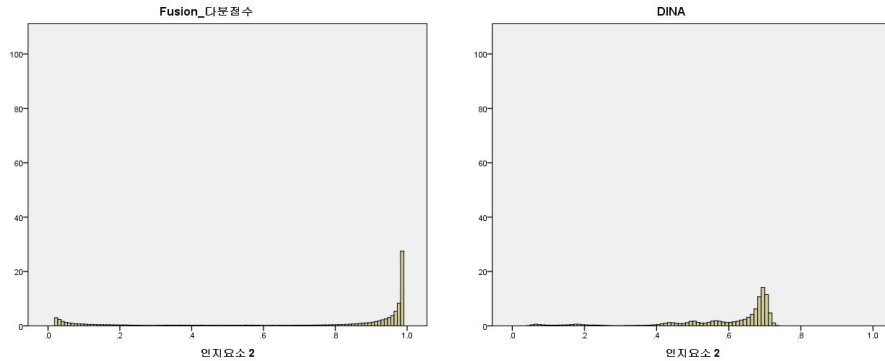
인지요소	Fusion_다분모형			DINA		
	상	중	하	상	중	하
1	36.76	21.97	41.28	56.72	29.50	13.78
2	66.09	16.13	17.78	0.00	93.56	6.44
3	51.35	37.31	11.34	47.01	42.94	10.05
4	72.09	8.61	19.30	70.96	11.81	17.22
5	69.88	23.63	6.49	0.00	100.00	0.00
6	69.42	26.38	4.20	93.11	0.00	6.89
7	40.85	28.04	31.11	53.97	34.09	11.95
8	68.17	18.00	13.83	41.81	52.06	6.13
9	69.42	23.55	7.03	44.89	52.54	2.57
10	77.46	7.02	15.52	84.10	0.00	15.90
11	61.14	21.68	17.18	52.92	32.23	14.84
12	70.85	21.55	7.61	0.00	89.44	10.56

〈표 8〉은 초6 수학 자료를 분석한 결과 산출된 PPM의 수치가 0.2 미만은 ‘하’, 0.2 이상~0.8 미만은 ‘중’, 0.8 이상은 ‘상’으로 분류한 결과를 Fusion Model과 DINA Model의 경우를 비교한 것이다. Fusion Model을 적용하였을 때, 각 인지요소를 ‘상’의 수준으로 습득하였다고 분류된 비율은 36.8~72.1%로 2011년 초6 수학과 학업성취도 평가에서 ‘우수학력’으로 구분된 학생 비율(21.4%)보다 높았다. 하지만 한 가지 특정 인지요소에서 우수한 것이 전체적으로 우수한 것보다 쉽기 때문에 수학과 전체 우수 비율인 21.4%보다 개별 인지요소에서 우수한 비율이 높은 것은 자연스러운 것으로 보인다. 반면 2011년 학업성취도 평가의 초6 학생들의 수학과 ‘기초학력 미달’ 비율은 1.0%였는데, Fusion Model에서 산출된 각 인지요소에서의 ‘하’ 비율은 4.20~41.28%로 나타났다.

〈표 8〉에서 음영으로 표시된 부분은 특정 수준에 해당하는 학생의 비율이 0%로 나타난 경우이다. DINA 모형을 적용하였을 때, 인지요소 6과 인지요소 10은 ‘중’ 수준의 학생이 0%인 것으로 나타났고, 인지요소 2와 인지요소 12는 ‘상’ 수준의 학생이 0%로 나타났다. 인지요소 5는 ‘상’과 ‘하’ 수준의 학생이 모두 0%로 나타났다.



[그림 3] 초6 수학 인지요소 5의 PPM 분포

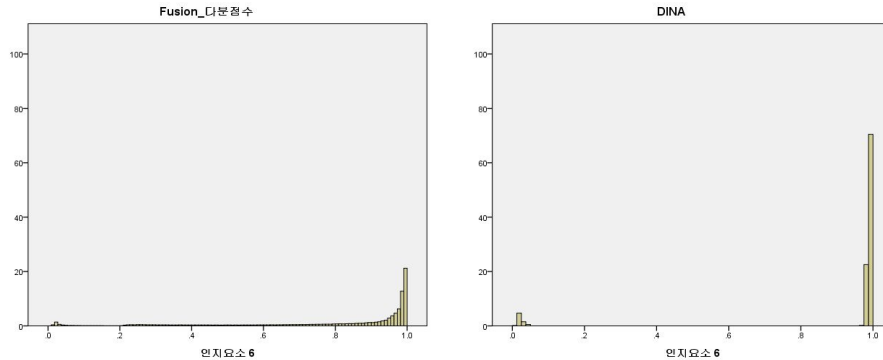


[그림 4] 초6 수학 인지요소 2의 PPM 분포

인지요소 5에서 DINA Model을 적용한 경우 ‘상’과 ‘하’ 수준의 학생이 0%로 산출되었기 때문에 [그림 3]에서는 두 가지 모형을 적용하였을 때의 PPM 분포를 제시하였다. 국어과의 경우와 마찬가지로 Fusion Model에서는 PPM이 넓은 점수대에 분포하고 있는 것과 달리 DINA Model의 경우 PPM이 중간 점수대 부분에만 분포하고 있는 것을 볼 수 있다.

또한, 인지요소 2에서는 DINA Model을 적용한 경우 ‘상’ 수준의 학생이 0%로 산출되었기 때문에 두 가지 모형을 적용하였을 때의 PPM 분포를 [그림 4]에 제시하였다. DINA Model의 경우 PPM이 0.8 이상에는 분포하지 않는 것을 볼 수 있다.

인지요소 6에서 DINA Model을 적용한 경우 ‘중’ 수준의 학생이 0%로 산출되었기 때문에 [그림 5]에서는 두 가지 모형을 적용하였을 때의 PPM 분포를 제시하였다. Fusion Model에서는 PPM이 넓은 점수대에 분포하고 있는 것과 달리 DINA Model의 경우 PPM이 0과 높은 점수대 부분에만 극단적으로 분포하고 있는 것을 볼 수 있다.



〔그림 5〕 초6 수학 인지요소 6의 PPM 분포

국어과의 결과에서와 마찬가지로 수학과 결과에서도 DINA Model의 경우에는 PPM의 분포가 편중되는 경향이 있는 것으로 나타났다. 이는 학업성취도 평가 자료의 인지요소별 프로파일을 숙달/미숙달 이상의 여러 단계로 구분하는 경우에 DINA Model이 한계점이 있음을 시사한다.

IV. 결론 및 논의

본 연구에서는 인지진단모형을 적용하여 국가수준 학업성취도 평가 결과의 활용성을 높이기 위한 방안을 찾고자 자료 특성에 가장 적합한 모형을 탐색하였다. 특히 지금까지 전수자료를 대상으로 인지진단모형을 적용한 연구 사례가 국내외를 통틀어 전무하기 때문에 자료의 크기가 60만명 이상인 대규모 학업성취도 평가 전수 자료 처리가 가능한 인지진단모형을 찾고자 하였다. 선행연구를 통해 대규모 평가 분석에 자주 활용된 인지진단모형은 Fusion Model과 DINA Model임을 파악하였고, 두 가지 모형 중에서 학업성취도 평가 자료에 더 적합한 모형을 탐색하고자 하였다.

두 가지 인지진단모형을 비교한 연구 결과는 다음과 같이 요약된다.

첫째, 이 연구에서는 60만명 이상의 대규모 평가 자료에 인지진단모형 분석을 위한 프로그램을 적용하여 개별 학생의 세부 인지요소 숙달 수준에 대한 프로파일 정보를 산출할 수 있는 가능성을 실증적으로 확인하였다. Fusion Model 또는 DINA Model을 적용할 때 분석용 프로그램이 개별적인 학업성취 프로파일을 산출하는 데 한 교과목당 2~3일이 소요되는 것으로 나타났다. 따라서 3개 학교급별로 여러 교과에 대한 분석이 필요한 학업성취도 평가의 경우, 초등

학교 3개 교과(국어, 수학, 영어), 중학교 5개 교과(국어, 사회, 수학, 과학, 영어), 고등학교 3개 교과(국어, 수학, 영어)를 분석하는데 11가지 자료 \times 3일 = 33일이 소요된다는 결론이다.

둘째, Fusion Model과 DINA Model을 비교한 결과 인지요소 숙달수준을 숙달/미숙달의 2단계가 아닌 상/중/하의 3단계로 분류하는 경우, DINA 모형은 '상'으로 분류되는 학생이 0% 또는 '하'로 분류되는 학생이 0%에 가깝게 나타나는 등 인지요소 숙달 수준이 편중되는 경향을 보였다. 따라서 국가수준 학업성취도 평가 자료를 인지진단모형을 통해 분석해 인지요소 숙달 수준을 2단계 이상의 더 많은 단계로 분류하는 경우 Fusion 모형이 더 적합한 것으로 나타났다.

인지요소 숙달 수준을 '상/중/하'와 같이 3단계로 구분했을 때의 장점은 분류일치도가 높아진다는 것이다. Roussos et al.(2007)은 인지진단모형 분석 결과를 0.2와 0.8을 기준으로, 즉 '0.2 미만 = 상', '0.2 이상~0.8 미만 = 중', '0.8 이상 = 상'으로 분류하였을 때, Cohen's kappa가 0.6 이상으로 유지되는 최적의 분류일치도가 나타났다고 보고하였다. 이러한 (0.2, 0.8) 기준은 현재 학업성취도 평가에서 보고하고 있는 4단계 성취수준을 구분할 때, '우수학력'과 '기초학력'을 분할하는 기준과 동일하다는 장점이 있다. 학업성취도 평가에서는 교과별로 교육과정 목표를 도달한 정도가 80% 이상인 경우에 우수학력, 50~80%를 보통학력, 20~50%를 기초학력, 20% 미만을 기초학력 미달로 규정하고 있으므로 각 교과별로 측정하고 있는 세부 인지요소에 대한 숙달 수준도 (0.2, 0.8)을 기준으로 3단계로 구분해주는 것이 더 자세한 정보를 제공해줄 뿐 아니라 학생, 학부모, 교사가 이해하기도 쉽다. 따라서 인지요소 숙달 확률 추정값이 편중되는 현상을 보이지 않아 상/중/하의 3단계로 고르게 분류되는 Fusion 모형이 국가수준 학업성취도 평가 자료 분석에 더 적합하다고 할 수 있다.

상/중/하의 3단계로 숙달 수준을 구체적으로 분류할 경우에 한 가지 더 장점이 있는데, 진단 결과에 따른 학습 계획을 수립할 시에도 매우 효과적일 수 있다. 예를 들어 <표 9>에서 학생 A와 B의 국어 성취수준은 우수학력으로 동일하다. 그러나 학생 A는 인지요소 1, 4, 8을 숙달하지 못하였고, 학생 B는 인지요소 8과 9를 숙달하지 못하였다. 이를 통해 학생들은 자신에게 부족한 국어 관련 지식과 기능에 대해 알 수 있다.

또한, 인지요소 숙달 단계를 상/중/하의 3단계로 구분해보면, 학생 A가 부족하다고 나타난 인지요소 1, 4, 8 중에서도 특히 인지요소 4의 숙달 수준이 '하'로 나타났다. 따라서 인지요소 4(세부 정보 파악)에 대한 학습이 가장 시급히 이루어질 필요가 있다. 학생 B의 경우에도 숙달/미숙달의 2단계로 판단하였을 때에는 인지요소 8과 9만을 숙달하지 못한 것으로 나타났는데, 3단계로 구분하였을 때에는 인지요소 1, 8, 9에서 '중'으로 나타나 인지요소 1을 포함한 세 가지 인지요소에 대한 학습이 상대적으로 중요한 것을 알 수 있다.

〈표 9〉 3단계 인지숙달 수준을 적용한 학생 학업성취 프로파일 예시(고2 국어)

성취 수준	학생	인지요소													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
우수 학력	A	X	O	O	X	O	O	O	X	O	O	O	O	O	O
		중	상	상	하	중	상	상	중	상	상	상	중	상	상
	B	O	O	O	O	O	O	O	X	X	O	O	O	O	O
		중	상	상	상	상	상	상	중	중	상	상	상	상	상
보통 학력	C	X	X	O	X	X	X	O	X	O	O	X	O	X	O
		하	하	상	중	하	하	중	중	상	상	중	상	중	상
	D	O	O	O	X	X	X	X	O	X	O	O	O	O	O
		상	상	상	중	중	하	하	상	중	상	상	중	상	중

이렇게 인지숙달 수준을 2단계 이상 상/중/하로 나누어 학생들에게 학업성취 결과를 제공할 경우, 학생들은 자신의 인지요소별 숙달 수준에 대한 정보를 토대로 학습 중요도에 따른 학습 순서를 계획할 수도 있다. 학생들은 숙달 수준이 '하'로 나타난 국어 인지요소에 대한 학습을 우선적으로 수행하되, '중'인 인지요소에 대한 추가적인 학습도 수행할 필요가 있다.

또한, 후속 연구로서 이러한 인지요소 숙달 수준 분류를 어떻게 할 것인지에 대한 수준설정(standard-setting) 방법에 대한 연구가 필요할 것으로 보인다. 국가수준 학업성취도 평가 결과 분석에 실제적으로 인지진단모형을 적용하여 학생들의 학업성취 프로파일 정보를 제공해주기 위해서는 각 인지요소에서의 숙달 수준을 어떠한 기준을 사용하여 분할할 것인지, 즉 수준 설정에 대한 구체적인 논의가 이루어져야 할 것이다. 물론, 숙달 수준을 숙달/미숙달, 상/중/하, 또는 그 이상의 단계로 구분하는 가능성에 대해서도 논의가 필요하다.

참 고 문 헌

- 김성훈(2005). 인지구조모형에 근거한 학생의 지식상태 진단. *교육학연구*, 43(1), 81-107.
- 김성훈, 송미영(2011). DINA 모형을 활용한 대규모 학업성취도 평가 결과 분석. *교육평가연구*, 14(1), 177-200.
- 김수진, 송미영, 김선희(2008). Fusion Model에 의한 수학 능력 진단을 위한 Q행렬의 정교화. *교육평가연구*, 21(2), 115-139.
- 김희경, 한정아, 최숙기, 김부미(2012). 인지진단모형을 적용한 학업성취 프로파일 분석 및 결과 보고 방안. 한국교육과정평가원 연구보고서 RRE-2012-7.
- 송미영, 이영선, 박윤수(2011). 인지진단모형을 통한 국가수준 학업성취도 평가 결과 분석 및 성적 보고 방법 탐색. 한국교육과정평가원 연구보고서. RRE-2011-8.
- Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, 15(2), 119 -57.
- Chiu, C., Seo, M., & Douglas, J. (2011). Cluster analysis for cognitive diagnosis: An application to the 2001 PIRLS reading assessment. *IERI Monograph Series 4*.
- College Board(2012). <http://www.collegeboard.com/student/testing/psat/scores/report.html>.
- de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45, 343-362.
- de la Torre, J. (2009). DINA model parameter estimation: A didactic. *Journal of Educational Behavioral Statistics*, 34, 115-130.
- Dibello, L., & Stout, W. (2010). *Arpeggio version 3.1 [Computer Program]*. Chicago: Applied Informative Assessment Research Enterprises.
- Dibello, L., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment 361-389*. Hillsdale, NJ: Erlbaum.
- Doornik, J. A. (2003). *Object-oriented matrix programming using Ox [Computer Program]*. London: Timberlake Consultants Press and Oxford.
- Fu, J. (2005). *A polytomous extension of the Fusion Model and its Bayesian parameter estimation*. Unpublished doctoral dissertation. University of Wisconsin- Madison.

- Fu, J., & Li, Y. (2007, April). *Cognitively diagnostic psychometric models: An integrative review*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Haertel, E. H. (1990). Continuous and discrete structure models to map skill structure of achievement items. *Psychometrika*, 55, 477-494.
- Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.
- Hennig, C. (2004). *Cluster-wise assessment of cluster stability*. Unpublished research report, Department of Statistical Science, University of College London.
- Henson, R. A., & Templin, J. L. (2004). *Modifications of the Arpeggio algorithm to permit analysis of NAEP*. Unpublished ETS project report, Princeton, NJ.
- Huebner, A., Wang, B., & Lee, S. (2009, June). *Practical issues concerning the application of the DINA model to CAT data*. Paper presented at 2009 GMAC conference on Computerized Adaptive Testing, Minneapolis, MN.
- Jang, E. E. (2009). Demystifying a Q-Matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, 6, 210 - 238.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258-272.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Lee, Y-W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening. *Assessment: Language Assessment Quarterly*, 6(3), 239-263.
- Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spain fellow working papers in second or foreign language assessment*, 9, 17-46.
- Park, C. (2008). *A multilevel IRT model for group-level diagnostic assessment with application to TIMSS*. Unpublished doctoral dissertation. University of Wisconsin-Madison.
- Roussos, L. A., Dibello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. H. (2007). The Fusion Model skill diagnosis system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education* 275-318. New York: Cambridge University Press.

- Rupp, A. A., & Templin, J. L. (2008). The effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement*, 68, 78-96.
- Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS* (Version 1.4) [Computer program]. Cambridge, UK: MRC Biostatistics Unit.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.
- Tatsuoka, K. K. (1985). A probabilistic model for diagnosing the misconceptions by the pattern classification approach. *Journal of Educational Statistics*, 10(1), 55-73.

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

ABSTRACT

An Exploration of the Appropriate Cognitive Diagnostic Model for NAEA(National Assessment of Educational Achievement): A Comparison between Fusion and DINA Models

Hee-Kyoung Kim

(Research Fellow, Korea Institute for Curriculum and Evaluation)

Chan-Ho Park

(Associate Research Fellow, Korea Institute for Curriculum and Evaluation)

This study sought the way of providing assessment results as the type of elaborate profile by applying cognitive diagnostic model to NAEA(National Assessment of Educational Assessment). Because any cognitive diagnostic study applied to census data is unprecedented in Korea and abroad so far, it is aimed to find cognitive diagnostic model that is available for processing census data of a large-scale assessment whose size of data is more than 600,000 people. Prior studies identified that Fusion model and DINA model were often used for analyzing a large-scale assessment and there needs to search for more appropriate model for NAEA data among two models. As a result of comparing Fusion model and DINA model, if mastery probability of cognitive attribute is classified into three levels of high, middle and low, not two levels of mastery and non-mastery, DINA model showed that mastery probability of cognitive attribute tended to be unequally distributed such that 'high' or 'low' is close to 0%. Therefore, if NAEA data is analyzed through cognitive diagnostic model and mastery probability of cognitive attributes is classified into the mastery levels more than two, Fusion model is revealed as more appropriate.

Key Words : National Assessment of Educational Achievement, Cognitive Diagnostic Model, Fusion Model, DINA Model, Q-matrix