https://doi.org/10.24303/lakdoi.2023.31.4.83

베이지언 네트워크에 기반한 인지진단모형의 영어 읽기 시험에 대한 적용*

최세일
(전남대학교)

Choi, Sae il. (2023). Application of bayesian network-based cognitive diagnostic modeling to small sample English reading comprehension test data. *The Linguistic Association of Korea Journal*, *31*(4), 83-111. Cognitive diagnostic models (CDMs), a family of classification models developed to provide fine-grained diagnostic information for learning and teaching in education, have increasingly been used in language testing. However, most of the previous CDM studies in language testing have mainly been conducted based on large samples from professional testing agencies. This trend makes it difficult for practitioners to apply the models in classroom assessment contexts for which the models were originally developed. Realizing this limitation, researchers working in CDMs have recently begun to turn their attention to the conditions in which CDMs can work for classroom assessments, especially with small sample sizes. Bayesian networks (BN) provide an efficient and intuitive framework for modeling complex systems of observable or latent variables and have been extensively employed in the data science as well as in intelligent tutoring systems for modeling students' learning progress. The framework has also huge potential to be well suited for diagnostic modeling of students' learning in classroom contexts. This study was to examine whether BN can be applied in cognitive diagnostic modeling for classroom assessments. After constructing a set of small test data (N=100, 150, 200) from a large real test, the study applied a BN-based CDM model to the data sets and compared with conventional CDMs its item parameter and attribute classification recovery. The results show that the BN-based CDMs yielded uniformly better estimates in all testing conditions than the conventional methods. The study then discusses its implications for the CDM applications in language testing.

주제어(Key Words): 진단정보(diagnostic information), 인지진단모형(cognitive diagnostic models), 문항모수(itemparameters), 분류정확도(classification accuracy), 베이지언 네트워크(bayesian networks)

* 이 논문은 2022년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2022S1A5B5A16055981).

1. 머리말

총합평가(summative testing)의 한계를 극복하기 위한 다양한 시도가 교육평가에서 이루어지고 있으며 인지진단모형(cognitive diagnostic models; 이하 CDM)이 하나의 대안으로서 많은 관심을 받고 있다(Lee & Sawaki, 2009; Rupp et al., 2010). 문항반응이론과 달리 CDM은 학생을 시험이 측정하는 다양한 인지 영역의 프로파일별로 분류하여 각 프로파일의 특성에 맞는 피드백을 제공함으로써 교수·학습의 개선을 목적으로 한다(Leighton & Gierl, 2007). 이러한 CDM의 특성은 학생의 시험자료에 대해 보다 세분화된 환류효과(feedback effect)나 성적 보고(score reporting)를 원하는 학생과 학부모의 요구에 부응할 수 있고, 형성평가의 기능을 강화하려는 현재의 교육정책과도 상응하여 영어평가 분야에서도 CDM에 관한 상당한 연구를 찾아 볼 수 있다(Li, 2011). 연구의 대부분은 주로 읽기·듣기 시험자료를 바탕으로 CDM 모형의 적용 가능성을 탐구하고 있으며(Buck & Tatsuoka, 1998; Harding et al., 2015; Kim, 2015; Lee & Sawaki, 2009; Li, 2011; Li et al., 2016; Li et al., 2021), 쓰기와 말하기 또는 문법에 관한 소수 사례도 찾아 볼 수 있다(Kim, 2011; Yi, 2017). 또한 측정학적 관점에서 시험 데이터에 대한 최적모형 선택 문제를 다루거나(Dong et al., 2021; Liu & Bian, 2021; Yi, 2017), CDM 절차에서 핵심적인 Q-행렬의 구성과 관련된 문제들을 조사하는 소수 연구도 볼 수 있다(DeCarlo, 2012; Jang, 2009; Li et al., 2021; Sawaki et al., 2009).

CDM에 관한 연구가 한 세대 넘게 진행됨에 따라 현재까지의 CDM 관련 연구에 대한 반성과 새로운 대안을 제시하는 연구 결과가 제시되고 있다(Chen & de la Torre, 2014; Ravand, 2016). 2009년 이후 교육평가 관련 저널에 발표된 36개의 CDM 적용 연구를 검토한 Sessoms & Henson(2018)에 따르면 대부분의 연구는 4~23개의 인지요인 (평균=8, 표준편차=4.96)을 1000명 이상의 샘플에 적용하고 있다. 이는 현재의 CDM이 기본적으로 국가·국제 단위 또는 평가 기관의 대규모 시험에 근거하고 있음을 의미하며, 소규모 학교 단위에서 시행되는 인지진단모형과는 상당한 괴리가 있음을 시사한다. 이에 따라 최근에는 CDM을 학생과 교사의 실제 교수·학습 상황에 보다 더 근접시키려 다양한 시도가 이루어지고 있다. 예를 들면, 기존의 일회성 진단평가에서 탈피하여 장기간에 걸친 학습 과정에서 CDM을 적용하려는 일부 연구를 볼 수 있으며(Lin et al., 2020; Pan & Zhan, 2020), CDM의 문항수를 최소화하고 문항 이용의 극대화를 위해 CDM을 컴퓨터 적응검사로 연계하려는 연구도 볼 수 있다(Finkelman et al., 2014; Li et al., 2023; Sorrel et al., 2021; Sun et al., 2021). 또한 CDM 시행 상황에서 교사가 학생 또는 시험 문항에 대해 갖고 있는 사전정보(pror information)를 모수 추정에 적용함으로써 소규모 샘플에 기초한 CDM 적용의 가능성을 탐구하는 베이지언(bayesian) CDM에 관한 관심이 점증하고 있다(Akabay & de la Torre, 2020; Sinharay & Almond, 2007; Wang & almond, 2019; Zhan et al., 2019).

한편 21세기 이후 교육·심리학의 양적 연구 방법론 분야에서는 데이터 과학(data science)으로 부터 변수들의 네트워크를 그래프이론(graph theory)을 이용하여 모델링하는 기법을 도입하여 기존의 방법론을 대체하려는 시도가 이루어지고 있다(Almond et al., 2015; Epskamp et al., 2018). 이 가운데 베이지언 네트워크(bayesian network; 이하 BN)는 체계를 이루는 변수들의 관계를 그래프이론에 따라 소규모 기본 단위로 분해한 후, 이 새로운 기본 단위에서 변수 사이의 조건부 독립관계를 이용하여 새로운 정보 전달 체계를 구축하고, 이후 새로운 정보가 유입되었을 때 이를 베이즈법칙(Bayes Theorem)을 이용하여 체계 전체를 업데이트하는 모델링 방식이다(Neapolitan, 2023). CDM과 관련하여 최근 Almond et al.(2015), Levy & Mislevy(2016) 등은 BN의 CDM 적용 가능성을 제시하고 있다. 이 새로운 체계는 기존의 CDM을 포괄하면서도 그래프를 이용하여 인지모형을 직관적으로 구성할 수 있는 장점 외에도 많은 다른 가능성을 내포하고 있는데 그 몇 가지를 살펴보면 다음과 같다. 첫째, 학생의 잠재적 인지요인에 대한 확률모형이 가능하며 둘째, 시험에 포함되지 않았으나 관련된 새로운 인지영역에 대한 예측이 가능하고 셋째, 모형이 모듈(module) 형태이므로 모형의 확장과 변형이 매우 용이하며, 넷째 모수 추정을 국소적으로 수행함으로써 모수 추정의 계산량(computational load)을 현저히 감소시킬 수 있다. 베이지언 네트워크에 기반한 CDM(bayesian network-based CDM; 이하 BN-CDM)의 이와 같은 특성은 전통적인 CDM이 적용되기 어려운 수 백명 내외의 학생들이 위계를 이루며 심화·확장되는 교과 과정을 배우는 우리나라의 학습 상황에서 CDM을 수행하기에 매우 유용한 기제를 제공할 가능성을 갖고 있다.

이 연구의 목적은 BN의 위와 같은 특성을 이용하여 BN-CDM의 학교 단위 소규모 영어 시험 데이터에 대한 적용 가능성을 탐구하는데 있다. 이를 위하여 이 연구에서는 실제 영어 읽기 시험 데이터로부터 소규모 샘플을 무선 추출하여 기존의 CDM과 BN-CDM을 적용한 후 그 결과를 상호 비교함으로써 BN-CDM의 적용 가능성을 측정학적 측면에서 조사하고, 이후 그 결과가 갖는 의미를 영어교육 평가의 인지진단모형 적용에 관하여 논의한다.

2. 선행연구 및 베이지언 네트워크 기반 인지진단모형

2.1. 영어 읽기 능력 구성 요소

영어학습자에게 전반적으로 통용되는 영어 읽기 능력이 어떤 요인으로 어떻게 구성되었는지에 대해서 아직까지 불분명하다. 그 이유는 영어 읽기에 대한 인지적 관점에 따라 구성요인을 다르게 설정하며, 설정된 구성요인이 학습자의 영어 능력에 따라 변화하고, 각 시험마다 서로 다른 인지요인을 선택하고 있기 때문이다. 예를 들면, Hughes(2003)은 교실 수

업 환경에서 교사가 평가를 위해 고려할 수 있는 영어 읽기 능력 수십 가지를 제안하고 있으나, 이는 주로 영어학습 교재의 직관적 분석에 기초한 것으로 실증적 데이터를 이용한 검증이 필요하다. 이와 대조적으로 Harding et al.(2015)는 L1, L2 영어 읽기 관련 문헌 조사와 실제 시험 데이터 분석에서 도출한 다양한 읽기 능력 구성 요인을 제안하고 있다. 영어 읽기 구성 요인에 대해서는 읽기 능력을 다양한 세부 요인의 조합으로 바라보는 구성주의(componential approach)와 하나의 통합된 단일 개념(unitary approach)으로 보는 견해가 대립되고 있다. 구성주의 견해는 주로 80~90년대 회귀분석 등의 통계적 분석을 통해 수립된 것으로(예: Davey, 1988) 개별 연구마다 서로 다른 구성 요인을 제안하는 등의 문제점이 있지만, 교재나 시험 또는 영어 교육 과정 개발에 주로 이용되고 있다(Grabe, 2009; Lumley, 1993). 반면 읽기 능력을 통합된 하나의 개념으로 보는 견해에서는 실제 시험자료를 이용한 실증적 분석에서 읽기 능력 구성 요인을 찾을 수 없으며(Alderson, 1990a, 1990b), 설령 구성 요인이 존재한다 하더라도 이를 일관되게 분류하는 것이 불가능하다고 주장하고 있다(Alderson & Lukmani, 1989; Carver, 1992).

한편 이러한 구성 요인의 상호 작용과 위계 구조에 대해서는 다양한 모형이 제시되고 있다. 읽기 구성 요인의 상호 작용과 관련하여 각 구성 요인의 상보적 관계(compensatory relation)를 지지하는 의견이 다수이나(Bernhardt, 2005; Bolt & Lall, 2003; Grabe, 2009; Stanovich, 1980; Uso-Juan, 2006), 개별 구성 요인의 독립적인 역할을 강조하는 시각도 있다(Gough & Tunmer, 1986; Hoover & Gough, 1990). 그러나 영어 읽기 영역의 CDM 적용에 있어서는 비상보적 모형이 압도적으로 다수이다(Buck et al., 1998; Jang, 2009; Jang et al., 2013; Kasai, 1997; Kim, 2014; Lee & Sawaki, 2009). 읽기 능력의 구성 요인이 어떤 위계 구조를 갖는지에 대해서도 일관된 합의를 찾기 어렵다(Ravand, 2020). 예를 들어, Alderson(1990a, 1990b)과 Alderson & Lukmani(1989)는 어휘, 문법, 문장 독해, 문단, 텍스트 전체 이해와 같은 순으로 읽기 능력이 어떤 위계를 구성할 것이라는 것이 일반적 통념이지만 이를 뒷받침하는 어떤 실증적 근거를 찾기 어렵다고 하였다. 이에 대해 Weir et al.(1990)는 읽기 전략과 같은 상위 능력은 어휘, 문법, 문장 단위 해독과 같은 하위 능력을 전제한다고 반론하였다. 그러나 Alderson(2000)는 하위 능력은 수험자의 영어 능력이 향상됨에 따라 상위 능력에 통합되어 분리하기 어렵다고 하였으며, 실증적 시험 자료(DIALANG) 분석에서도 읽기 능력의 어떤 특정 요인이 읽기 수준을 구분한다는 근거를 찾을 수 없다고 하였다.

구성 요인과 관련된 또 다른 문제는 읽기에서 어휘와 문법 능력의 상대적 역할이다. 읽기에서 어휘 능력의 역할에 대해서는 많은 연구에서 그 중요성을 확인하고 있다(Nation, 2001; 2006; Morvay, 2012). 심지어 일부 연구에서는 텍스트를 빠르고 정확하게 읽기 위해서는 해당 텍스트 어휘의 95% 또는 98% 이상을 알고 있어야 한다고 주장하고 있다(Hu & Nation, 2000; Laufer, 1989; Nation, 2006). 그러나 읽기에서 문법 능력의 역할에 대해서는

다양한 견해를 보이고 있다. 일부 연구에서는 문법 능력이 상당한 비중을 차지한다고 보고하고 있으나(Afflerback et al., 2013; Gottardo & Mueller, 2009; Shiotsu & Weir, 2007), 다른 연구에서는 어휘 능력을 고려한 이후에 문법 능력은 큰 비중이 없는 것으로 간주하고 있다(Morvay, 2012; Lee, 2016; Van Gelderen et al., 2003; Zhang, 2012). 서로 상반되는 이러한 견해에 대하여 Choi and Zhang(2021)는 문헌 고찰(systematic review)을 통해 연구 방법, 샘플 크기, 실험 참가자의 능력 분포 등과 같은 실험 조건에서 큰 차이가 있어 어휘와 문법의 상대적 역할을 정확히 규정할 수 없다고 설명하고 있다.

2.2. 영어 읽기 능력에 대한 인지진단 모형 적용

영어 평가에서 CDM 적용은 주로 읽기 분야에 집중되어 있으나 실증적 연구 사례는 많지 않으며, 여기에서는 영어 평가 관련 주요 저널에서 볼 수 있는 읽기 시험 자료에 대한 CDM 분석 사례를 살펴본다. 초기 CDM 연구는 주로 대형 시험자료를 이용하여 CDM 적용 가능성을 조사하고 있는데, Buck et al.(1998)는 CDM 초기 모형인 규칙장모형(rule space model)을 일본의 토익응시자(N=5000) 시험 자료에 적용하여 분석하고 있다. 이 연구에 의하면 최종적으로 추출된 16개의 인지요인과 4개의 상호 작용이 읽기 성적 분산의 97%를 설명하고 약 91%의 분류정확도를 보였다고 한다. Buck et al.(2004)는 같은 규칙장모형을 SAT-Verbal 데이터에 적용하여 13개의 인지요인과 6개의 상호작용 모형이 읽기 성적 분산의 97%를 설명하고 90%의 분류정확도를 갖는다고 보고하였다. 초기의 소규모 규칙장 이론 적용 이후에는 주로 Fusion모형(Hartz & Roussos, 2005)을 적용하였는데, Jang(2009)는 LanguEDGE 시험 데이터(N=2700)에 Fuson 모형을 적용하고 이후 비진단용 시험데이터에 CDM을 적용하는 과정에서 발생하는 타당도 검증 문제를 다루고 있다. Lee & Sawaki(2009)는 TOEFL iBT 시험데이터(N=2720)에 Fusion 모형과 서로 다른 두 CDM 모형(General Diagnostic Model: von Davier & Yamamoto, 2004; Latent Class Analysis: Maris, 1999)을 적용하여 그 결과를 비교하였다. 이 연구에 의하면 모형간 차이는 크지 않으며 인지프로파일의 분포가 양 극단이 큰 비중을 차지하는 U자 형태 분포(U-shaped distribution)를 보인다고 하였다. 영어 읽기 영역의 CDM에서는 비진단용 시험자료에 CDM모형 적용을 하는 경우가 많은데, Jang et al.(2013)은 케나다 지역 주정부에서 시행한 학력측정검사(N=120,767) 데이터와 설문조사 자료를 이용하여 아동들의 출신배경에 따른 읽기 능력의 차이를 CDM으로 분석하고 있다. 또한 Kim(2015)는 한 미국 대학의 ESL 배치고사(N=1982) 자료에 CDM을 적용하여 분석함으로써 추가적인 진단정보를 제공할 수 있음을 보여주고 있다.

위에서 살펴본 읽기 영역 CDM은 대부분 대규모 데이터를 이용하여 CDM 모형의 적용 가능성과 문제점을 논의하거나 비진단용 시험에서 진단정보를 추출하기 위해 CDM을 이용하고 있다(Ravand, 2020). 이는 CDM의 영어 평가 분야 도입 초기에 이론적 토대를 마련하

기 위한 불가피한 과정으로 생각되나 실제 학교 단위에서 실행할 수 있는 진단평가와는 상당한 괴리가 있음을 알 수 있다. 또한 연구 결과의 해석에도 상당한 주의가 요구된다. 예를 들면, 규칙장 모형을 적용한 초기 CDM의 경우(Buck et al., 1998; 2004) 인지요인의 수가 과도하게 많아, 인지요인이 16개일 경우 상호작용이 없는 단순 이항변수로 가정해도 65536개의 인지프로파일이 이론적으로 가능하여 모수 추정에 대규모 데이터가 필요할 뿐만 아니라 각 프로파일의 해석과 이에 대한 피드백이 사실상 불가능함을 알 수 있다. 또한 비정상적으로 높은 읽기 분산 설명력은 특정 모형을 데이터에 적용했다기 보다는 데이터에 모형을 맞춰가는 과정을 수없이 되풀이 한 결과임을 이해해야 한다. 이러한 탐색적 모형 적용(exploratory model fitting)은 이론 개발 초기에 필요한 과정으로 판단되며 이러한 연구의 가치는 동일한 시험에서 유출할 수 있는 인지요인의 최대치를 실증적으로 파악하는데 있다고 볼 수 있다. 한편 영어 평가의 CDM 관련 연구, 특히 초기 CDM의 경우 연구자들이 특정 통계 패키지와 특정 모형을 반복 사용함을 관찰할 수 있는데(Li et al., 2016), 이는 당시에 CDM 프로그램의 기능과 종류가 극히 제한적이서 Fusion 모형에 특화된 Arpegio라는 프로그램에 의존해야 했고, 이 프로그램의 기본 모형(default model)이 RUM(reduced unified model: DiBello et al., 1995)이었기 때문이다.

2.3. 교실 환경에서 CDM 적용을 위한 연구

많은 잠재적 장점에도 불구하고 CDM은 교수·학습을 위한 진단평가라는 본래 목적 외에 다른 용도로 이용되고 있다(Lee & Shin, 2020). 최근 CDM 분야에서는 교실 수업 환경에서 실행 가능한 진단평가를 위해 다양한 연구가 진행되고 있는데 주로 소규모 샘플 환경에서의 최적모형 선택, 적정 인지요인 수 등에 관한 연구가 진행되고 있다. Sorrel et al.(2021)은 소규모 샘플(N=100, 인지요인 수=5, 문항수=30) 환경에서 DINA 모형(de la Torre, 2009)을 적용한 최적 모수 추정 방안 탐색을 위한 시뮬레이션 연구를 수행하였다. 연구 결과에 따르면 이러한 소규모 샘플에서는 현재 CDM 적용에서 가장 일반적으로 쓰이는 모수 추정 방식(MMLE-EM)이 상당한 편차(bias)를 초래하므로 베이지언 추정방식과 같은 대안을 고려해야 한다고 하였다. 한편 Akbay(2020)은 소규모 샘플 환경에서 비모수적 추정 방식(nonparametric approach)이 베이지언 추정방식과 유사한 결과를 보인다고 하여 또 다른 대안을 제시하고 있다. Sen & Cohen(2021)은 다양한 CDM 적용 환경(샘플크기:50~5000, 문항수:12~36, 요인수:3~5, CDM 모형: DINA, DINO, C-RUM, LCDM)아래 대규모 시뮬레이션 연구를 시행하였다. 연구 결과에 따르면, 안정적인 모수 추정을 위해서는 샘플 크기가 500명 이상이어야 하며, DINA/DINO처럼 모형이 단순해야 하고, 적어도 15~20개의 문항수를 유지해야 하며, 요인 수는 3개 이하로 제한해야 한다고 하였다.

Najera et al.(2023)과 Sen & Cohen(2021)의 연구는 소규모 교실 수업 환경에서 CDM

적용을 위해 나아갈 방향을 제시하고 있다. 첫째, 인지모형의 요인 수를 줄여서 모형의 크기를 최소화해야 하여 최소한의 문항 수를 유지하도록 해야 한다. 둘째, 포화모형(saturated model)과 같은 과도하게 복잡한 인지모형보다는 단순한 모형을 적용하도록 해야 한다. 셋째, 모수 추정시 베이지언 추정 방식이나 비모수적 추정 방식 또는 기타 다른 방식을 채택하여 샘플 크기가 200보다 작은 상황에서도 모수 추정의 안정성을 확보할 수 있도록 해야 한다. 현재 CDM이 대규모 데이터를 필요로 하는 근본적인 이유는 인지요소의 모수 전체를 동시에 추정하려는 추정 방식에 그 문제가 있다. 이 연구에서 탐구하려는 베이지언 네트워크는 인지 요소간의 위계를 구성하여 모수 추정을 국소적으로 시행함으로 계산량을 현저지 감소시킬 수 있을 뿐만 아니라 베이즈법칙을 이용하여 모수를 업데이트하므로 교과전문가가 갖는 사전정보를 모수 추정에 도입할 수 있어 소규모 샘플에서 CDM을 적용할 수 있는 매우 효율적인 접근이라 볼 수 있다.

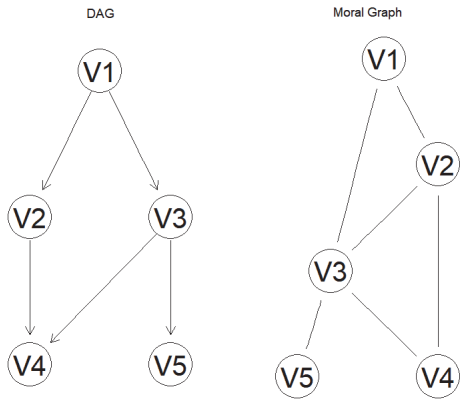

그림 1. DAG와 Moral Graph

2.4. 베이지언 네트워크 기본 개념

BN은 그래프 이론에 기초를 두고 있으며, 모든 그래프(G=<V, E>)는 노드(node, 또는 vertex)와 노드를 연결하는 선(edge 또는 link)으로 구성된다. 인지모형의 경우 노드는 인지요인이나 인지요인을 측정하는 문항을 나타내며, 노드 사이의 선은 이들의 관계를 나타낸다. 노드의 방향성 여부에 따라 방향성·비방향성 그래프(directed/undirected graph)로 구분되는데, BN에서는 언제나 방향성 그래프를 이용한다. '그림 1'은 5개의 노드(V1~V5)로 이루어진 방향성 그래프이다. 또한 BN에서는 원점회귀(cyclic) 그래프(예: V1->V2->V3->V1)는 허용되지 않아 방향성 비순환 그래프(direted acyclic graph; 이하 DAG)만을 다룬다. 노드의 관계는 친족 관계를 나타내는 명칭으로 기술되는데, '그림 1'에서 V1은 V2, V3의 부모 노드

(parent node)이며 V4는 V2와 V3의 자식 노드(child node)이다. 또한 V1은 V4와 V5의 조상 노드(ancestor node)이며, V4와 V5는 V1의 후손 노드(descendant node)이다.

BN에서는 어떤 체제(system)를 구성하는 수 많은 노드(변수)의 관계를 단지 세 개의 기본 관계로 분석하며, '그림 1'의 노드 관계를 이용하여 설명하면 다음과 같다. 첫 번째 관계는 체인 또는 시리얼 관계(chain/serial connection)이며 '그림 1'의 DAG에서 (V1->V2->V4), (V1->V3->V4), (V1->V3->V5)가 그 예이다. 이러한 체인 관계에서는 첫 번째 변수가 세 번째 변수에게 영향을 미치는데 반드시 중간 변수의 매개 과정을 거친다는 것이다. 예를 들면, (V1->V2->V4)의 경우 V1이 V4에게 영향을 미치나 반드시 V2 변수를 통해서 그 영향이 매개되며, BN 모형에서 V2의 값이 V4에 입력되면 V1은 V4와 조건부독립(conditional independence) 관계가 된다. 이제 이러한 관계를 세 변수의 확률분포 모형으로 기술하면 아래 식 (1)과 같다.

$$P(V1, V2, V4) = P(V1)P(V2|V1)P(V4|V2) \tag{1}$$

위 식(1)이 갖는 의미는 BN에서 매우 중요한데 그 이유는 다음과 같다. 식 (1)의 왼쪽 항은 세 변수의 결합확률분포(joint probability distribution)를 나타내며 이는 세 변수의 확률분포를 계산할 때 3차원 모수추정(3-dimensional estimation)과정을 거쳐야 함을 의미한다. 그러나 오른쪽 항은 모두 일차원 주변분포(P(V1))나 일차원 조건부분포(P(V2|V1), P(V4|V2)의 곱(product)으로 이루어져 계산량이 현저하게 감소됨을 의미한다. 만일 인지요인 이 16개로 구성된 모형의 확률 분포를 계산할 경우, 식 (1)의 좌변항을 이용하면 16차원 결합확률분포함수를 고려해야 하지만, 우변항을 따르면 1차원 주변분포와 조건부분포의 순차적 곱으로 계산할 수 있음을 의미한다. BN 모형에서 중요한 두 번째 관계는 발산 관계(divergent connection)이며 '그림 1'에서 (V2<-V1->V3), (V4<-V3->V5)이 그 예이다. 이 변수들의 관계를 (V2<-V1->V3)의 경우로 설명하면 다음과 같다. 변수 V2와 V3는 V1의 자식노드이므로 V1으로부터 공통의 영향을 받아 서로 독립적일 수 없다. 그러나 V1으로부터 받은 특성을 고려한 이후에 두 자식 노드는 서로 독립적이다. 이제 이러한 관계를 세 변수의 확률분포 모형으로 기술하면 아래 식 (2)와 같으며, 우변항이 갖는 계산상의 이점은 식 (1)과 같다.

$$P(V1, V2, V3) = P(V1)P(V2|V1)P(V3|V1) \tag{2}$$

BN 모형의 세 번째 기본 관계는 수렴관계(convergent connection)이며 '그림 1'에서 (V2->V4<-V3)이 그 예이다. 이 변수들의 관계를 BN 모형이 실제로 많이 쓰이는 유전학적 입장에서 설명하면 다음과 같다. 변수 V2(부)와 V3(모)는 V4의 부모이지만 유전적인 입장에서 자식 V4를 낳기 전까지 서로 독립적이다. 그러나 자식 V4가 어떤 유전적 특성(질환)을

발현하게 되면 부모 V2, V3는 유전적 입장에서 서로 독립적일 수가 없다. 왜냐하면 V4의 유전질환이 V2로 물려받음이 판명되면 V3는 귀책사유가 없으며 그 반대의 경우도 성립한다. 또 다른 경우에는 두 부모의 유전자가 상호결합하여 문제를 일으킬 수 있으므로 어떤 이유에서든지 V4의 상태를 확인한 이후에는 독립적일 수가 없다. 이 관계를 인지모형에 적용하면, 다수의 인지요인이 어떤 문항 해결에 연관될 경우 해당 문항에 대한 반응을 설명할 때 반드시 연관된 인지요인을 동시에 고려해야 함을 의미한다. 이제 이러한 관계를 세 변수의 확률분포 모형으로 기술하면 아래 식 (3)과 같으며, 우변항이 갖는 계산상의 이점은 식 (1)과 같다.

$$P(V2, V3, V4) = P(V2)\,P(V3)\,P(V4|V2, V3) \tag{3}$$

위 식 (1), (2), (3)의 관계를 일반화하면 아래 식 (4)로 나타낼 수 있다.

$$P(V_1, V_2, V_3, ,,, V_K) = \prod_{i=1}^{K} P(V_i|parents(V_i)) \tag{4}$$

위 식 (4)에서 $parents(V_i)$는 변수 V_i의 부모 변수이며 어떤 체계를 이루는 수 많은 변수(예: K=100, 300)의 결합확률분포를 식 (1), (2), (3)처럼 각 변수의 부모 변수에 대한 주변 또는 조건부 확률분포의 곱으로 변환할 수 있음을 의미한다.

2.5. 베이지언 네트워크에서 정보 업데이트 과정

위 식 (4)의 원리를 이용하여 다양한 DAG 모형에서 모형의 구조(structure)와 모수(parameter)를 어떻게 추정하는지는 매우 기술적인 문제로 다양한 문헌에 기술되어 있다(예: Neopolitan, 2004; Koller & Friedman, 2009). 이 연구에서는 직관적으로 이해하기 쉬운 Cowell et al.(2007)의 junction tree에 의한 베이지언 네트워크의 정보 전파(information propagation) 과정을 개념적으로 설명한다. '그림 1'과 같은 DAG가 확정되면 전체 네트워크를 계산이 용이하도록 부분집합(cliques)으로 분할 하는데, 여기서 주의해야 할 사항은 입력된 정보가 중복이나 누락 또는 단절 없이 네트워크 전체로 정확하게 전파되어야 한다는 것이다. 예를 들면, 식 (3)의 우변 마지막 항($P(V4|V2, V3)$)에서 변수 V4는 부모변수 V2와 V3에 근거하고 있음이 표현되어 있고, DAG를 소규모 그룹으로 분할할 때 이 내용이 정확히 반영되어야 한다. 이를 위해 DAG 분할 첫 단계에서는 다수의 부모 변수가 공통의 자식 변수에 영향을 미치는 경우 부모 변수를 동시에 고려할 수 있도록 서로 연결하는 과정(moralization)이 필요하다. '그림 1'의 오른쪽 moral graph를 보면 왼쪽 DAG에서 수렴 관

계를 보이는 (V2, V3, V4)에서 부모변수 (V2, V3)를 의도적으로 연결하여 묶어 두었음을 알 수 있다. 또한 부분집합의 확률관계 계산에서는 노드를 연결하는 화살표의 방향이 무의미하므로 이를 제거하였다. 두 번째 단계에서는 계산을 소규모 국소 단위로 수행할 수 있도록 변수가 4개 이상인 부분 집합에서는 임의의 연결선을 만들어 전체가 삼각형 구조가 되도록 만든다(triangulation). '그림 1'의 DAG에서 변수 (V1, V2, V3, V4)는 4개의 변수가 연결되어 있으므로 3개의 삼각형 구조로 분할해야 한다. 이 경우 두 가지 분할 방식이 가능한데 '그림 1'의 moral graph를 보면 수렴 관계의 부모변수인 (V2, V3)를 연결하여 두 개의 삼각형으로 분할하였음을 알 수 있다. 또 다른 방안으로 V1과 V4를 연결하여 삼각형을 만들 수도 있는데, 이 경우 V4에 영향을 미치는 부모변수 (V2, V3)를 서로 연결하지 못하는 문제가 발생하므로 대안이 될 수 없다. 세 번째 단계에서는 분할된 부분집합(clique)이 원래 네트워크의 모든 변수를 빠짐없이 포함하고 있는지 확인해야 한다. '그림 1'의 DAG는 세 개의 부분집합(cliques), C1=(V1, V2, V3), C2=(V2, V3, V4), C3=(V3, V5)로 분할 되었고 이 세 개의 부분집합이 전체 네트워크 노드를 빠짐없이 포함하고 있다. 마지막 단계에서는 분할된 부분집합 간의 연결부(running intersection)를 만들고 이를 통해 부분집합(cliques) 사이의 정보 전파를 시행한다. '그림 1'에서 C1=(V1, V2, V2)과 C2=(V2, V3, V4)의 경우 변수 (V2, V3)가 C1, C2의 공통변수이므로 P(V2, V3)가 연결부가 되며, C2=(V2, V3, V4)와 C3=(V3, V5)는 변수 (V3)가 공통변수이므로 아래 '그림 2'에서처럼 P(V3)가 연결부이다.

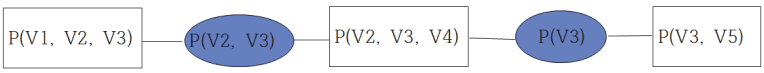

그림 2. Junction Tree와 Running Intersection

'그림 2'의 junction tree에서 실제로 어떻게 확률 계산이 이루어지는지 그 예가 '표 1'에 제시되어 있다. 모든 노드가 이항변수(예: High/Low, Mastery/Non-mastery, Pass/Fail 등)라고 가정하고 각 노드를 측정하는 문항이 한 개라고 가정한다('그림 1'의 DAG에는 인지요인 모형만 제시되어 있으며, 각 인지요인을 측정하는 문항은 공간 제약상 생략되어 있다). 이제 인지요인 V1을 측정하는 한 문항에 대해 어떤 학생이 정답으로 반응하였을 경우 이 정보가 '그림 2'의 P(V1, V2, V3) 부분집합에 전달되고, 이 부분집합에 상응하는 조건부확률분포표(conditional probability table; 이하 CPT) 테이블로부터('표 1'의 왼쪽 P(V1, V2, V3) 부분) P(V2, V3)가 계산되며, 이 정보는 다시 P(V2, V3, V4)에 전달되어 P(V4)에 대한 정보를 업데이트하게 된다. 나머지 부분집합 P(V2, V3, V4)와 P(V3, V5) 사이의 정보 전달도 같은 방식으로 작동되며 모든 변수가 네트워크로 연결되어 있으므로 정보 업데이트가 어느 부분집합부터 시작되는지에 관계없이 동일한 결과를 갖는다.

표 1. Conditional Probability Table(CPT)

P(V1, V2, V3)					P(V2, V3)				P(V2, V3, V4)			
V1	V2	V3(H)	V3(L)		V2	V3(H)	V3(L)		V2	V3	V4(H)	V4(L)
H	H	0.75	0.25	<=>	H	0.37	0.13	<=>	H	H	0.56	0.44
H	L	0.72	0.28		L	0.19	0.31		H	L	0.48	0.52
L	H	0.65	0.35						L	H	0.42	0.58
L	L	0.10	0.90						L	L	0.20	0.80

2.6. 문항반응과 인지요인의 결합 방식

인지요인은 개념상 연속변수일지라도 CDM에서는 범주형 변수(categorical variable: 상/중/하, 수/우/미/양/가, 높음/낮음, 완전학습/불완전 학습 등)로 처리하며, 인지요인을 측정하는 문항은 대부분 이항변수(정답/오답)이거나 다항변수(부분점수 0, 1, 2, 3)이다. BN-CDM에서는 문항에 대한 학생들의 반응을 입력받아 인지요인에 대한 확률분포를 모형화하기 위해 아래 식 (5)와 같은 문항반응이론(item response theory) 방정식을 이용한다(Almond et al., 2015).

$$P(X_i = 1|\theta_j) = \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]} \tag{5}$$

식 (5)를 이용하기 위해서는 몇 가지 선결 과제가 있다. 첫째, 식 (5)의 인지요인 (θ)은 연속변수인 반면(대부분 정규분포를 가정함) CDM에서는 범주형 변수이므로 연속변수 (θ)를 범주형 변수로 변환해야 한다. 다양한 접근법이 있지만 Almond et al.(2015)은 이 문제를 연속변수(θ)의 범위를 정규분포 범위(대략 $-3 < \theta < 3$)내에서 범주형 변수의 단계(level)에 따라 분할한 값으로 대체한다. 예를 들어 V1=θ1이 이항변수(높다/낮다)이면 V1=θ1(=높다/낮다)에 (0.67, -0.67)값을 각각 부여한다. 만일 V1=θ1이 3개의 범주값을 갖는다면(예: 상/중/하) V1=θ1(=상,중,하)에 (0.97, 0, -0.97)를 부여하며, 단계가 네 개의 경우 각각 (1.15, 0.31, -0.31, -1.15)값을 부여한다. 둘째, 식 (5)는 인지요인이 하나인 경우이나 BN-CDM을 비롯한 거의 모든 CDM은 다수의 인지요인을 상정하므로 이 다수의 인지요인이 어떻게 결합하는지 결정해야 한다. Almond et al.(2015)는 인지요인의 다중결합방식을 예시하고 있는데 주요 결합방식을 살펴보면 다음과 같다. 첫째, DINA 모형처럼 문제를 해결하는데 모든 요인이 필요한 경우(conjunctive model), 필요한 요인 중 최소값이 문항 반응에 결정적인 역할을 하므로 다음과 같은 결합방정식을 사용한다.

$$Z(\theta_1, \theta_2, ...\theta_K) = \text{다중요인 결합체} = MIN(\alpha_k\theta_k) - \beta \tag{6}$$

또한, 다수의 요인 중 가장 큰 값을 갖는 요인이 문항 반응을 결정하는 경우(disjunctive model) (예: CDM에서 DINO, NIDO, GDM 모형) 다음과 같은 결합방정식을 이용한다.

$$Z(\theta_1, \theta_2, ...\theta_K) = \text{다중요인 결합체} = MAX(\alpha_k\theta_k) - \beta \quad (7)$$

한편, 다수의 요인이 서로 상보적으로 문항 반응을 결정하는 경우(compensatory model) (예: CDM에서 ACDM, NC-RUM, GDINA) 각 인지요인 합의 평균값이 문항 반응을 결정하므로 다음과 같은 식으로 나타낸다.

$$Z(\theta_1, \theta_2, ...\theta_K) = \text{다중요인 결합체} = \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\alpha_k\theta_k - \beta \quad (8)$$

마지막으로 식 (5)를 이용하기 위해서는 각 문항의 난이도(β)와 변별도(α)값을 결정해 줘야 하는데 이는 문항 개발자나 연구자의 사전정보에 따라 적절한 값(예: β=(-1, 0, 1), α=(0, 0.5, 1))의 조합으로 사용한다.

3. 연구 방법

3.1. 시험 데이터와 연구 참가자

이 연구에서 사용한 시험자료는 한 종합대학의 어학연구소에서 자체 개발한 토익 모의고사 읽기 평가 자료이다. 대학생의 취업 역량 강화를 위해, 학교 당국은 토익으로 측정되는 어학 능력의 체계적 관리 중요성을 인식하고 토익성적 관리를 위한 프로그램을 기획하였다. 이 계획에 따라 재학생은 1년에 적어도 1회 이상 토익 모의고사를 의무적으로 응시해야 하였다. 해당 모의고사는 어학연구소가 구축한 기존의 문제은행으로부터 토익전담 연구원들에 의해 개발되었으며, 정규 토익과 같은 형식으로 듣기·읽기 각각 100문항으로 구성되었다. 개발된 토익 모의고사는 1학년 3018명, 2학년 1650명이 응시하였으며, 시험은 컴퓨터화 검사(computer-based test)로 각 학과 별로 정해진 날짜에 시행되었다.

이 연구는 토익시험 자체에 대한 CDM 적용이 아닌 영어 읽기평가 자료에 대한 BN-CDM의 적용 가능성을 탐색하기 위한 것이다. 이를 위해 CDM 목적에 부합하도록 소수 문항으로 시험을 재구성하되 토익 읽기 영역 전체와 비슷하도록 100문항 중 문법·어휘 영역 8문항, 1개 지문과 2개 지문 읽기 각각 7문항, 5문항(총 20문항)으로 시험을 구성하였다. BN-CDM 연구와 관련된 앞선 연구에서는 기존의 CDM과 BN-CDM 비교 연구를 위해

2학년 응시 데이터를 사용하였다. 이 연구에서는 같은 시험의 1학년 응시 데이터를 모집단 데이터(population data)로 설정하고 이 데이터로부터 크기가 다른 샘플(N=100, 150, 200)을 랜덤 추출하여 소규모 데이터에 기초한 BN-CDM의 측정학적 성격을 조사하였다.

3.2. Q-행렬 구성

CDM에서는 시험의 각 문항과 해당 문항이 측정하는 인지요인을 연결하는 Q-행렬 구성이 중요한 과제이다(Lee & Sawaki, 2009). 인지요인 구성을 위해 우선 일본 영어학습자를 대상으로 토익 읽기 영역(N=5000)에 대해 rule space 모형을 적용한 Buck et al.(1998)의 결과를 참고하였다. 또한 읽기 영역에 대한 CDM 선행연구(예: Kim, 2015; Lee & Sawaki, 2009b; Li, 2011; Li, Hunter, & Lei, 2016; Sawaki, Kim, & Gentile, 2009; Svetana et al., 2011)와 읽기 능력 평가 관련 문헌(예: Alderson, 2000; Hughes, 2003)을 병행하여 검토하였다. 한편 토익시험은 시험 구성에서 고유의 명시적 영역(어휘, 문법, 문단 읽기 등)을 갖추고 있는 점을 고려하여 최종적으로 읽기 인지영역을 문법, 단어, 주제 파악, 세부 사항, 추론으로 결정하였다. Q-행렬 구성에서 각 문항과 인지요인에 대한 선택은 저자 본인과 영어교육 박사과정을 이수하고 있는 7명의 대학원생이 먼저 개별적으로 검토하고, 다시 전체가 모여 합의를 통해 최종적으로 결정하였다. 이후 모집단 모수 추정 과정에서 GDINA 패키지(Ma & de la Torre, 2016)의 Q-행렬 수정 과정을 거쳤으며 시험 반응 데이터 생성에 이용된 최종 Q-행렬은 '표 2'에 제시되어 있다.

표 2. Q-행렬

문항	vocabulary	grammar	gist	details	inference
q01	1	0	1	0	0
q02	0	1	1	0	0
q03	0	1	1	0	0
q04	0	1	0	0	0
q05	0	1	0	0	0
q06	1	0	0	0	0
q07	1	0	1	1	0
q08	1	0	0	0	0
q09	0	0	1	1	0
q10	1	0	1	0	0
q11	0	0	1	1	0
q12	0	0	1	1	0
q13	0	0	0	1	0

문항	vocabulary	grammar	gist	details	inference
q14	0	0	1	1	0
q15	0	0	1	1	1
q16	0	0	1	1	0
q17	0	0	1	0	0
q18	0	0	1	1	1
q19	0	0	0	0	1
q20	1	0	1	1	0

3.3. BN-CDM 모형 구성

전통적인 CDM과 달리 BN-CDM은 인지요인의 위계관계(hierarchical relationship)를 모형에 명확히 도입해야 하는 특징이 있다. 이 연구에서는 전통적인 CDM처럼 위계를 고려하지 않은 모형(unstructured)과 '문법'(grammar)과 '어휘'(vocabulary)능력이 일반적인 읽기 능력 '주제파악'(gist), '세부정보 파악'(specific details), '추론'(inference) 능력에 대해 '그림 3'처럼 위계를 갖는 두 모형을 고려하였으며, 모집단 반응 데이터 생성은 위계모형을 이용하였다. 위계모형은 문법 능력이 읽기에 영향을 미치지만 그 효과는 어휘를 거쳐서 매개되는 모형을 의미한다. 이와 같은 모형 설정의 동기는 상당히 많은 선행 연구에서(예: Morvay, 2012; Nation, 2015; Susoy & Tanyer, 2019; Zhang, 2012) 어휘 능력의 상대적 중요성을 강조하고 있기 때문이다. '그림 3'에서 문법과 어휘의 위치가 뒤바뀐 모형도 고려할 수 있으나, 선행연구에서 현재의 모형이 훨씬 더 나은 모형적합도를 보였기 때문에 생략하였다.

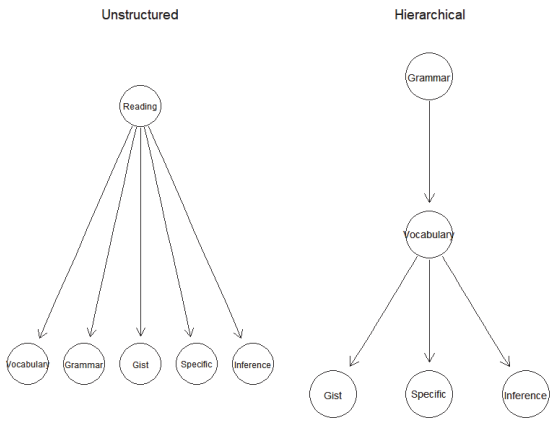

그림 3. BN-CDM모형(unstructured/hierarchical)

3.4. 모수 추정 과정과 결과 비교

토익 읽기 영역에 대한 1학년 데이터를 모집단으로 설정하였기 때문에 DINA모형의 문항모수(guessing, slip)와 인지프로파일 분포를 전통적인 CDM(DINA)모형과 BN-CDM을 적용하여 추정하였다. BN-CDM에서 모형의 추정은 이 모형 적용에 이용할 수 있는 프로그램(CPTtools, Rnetica; Almond et al, 2015)에서 연구자가 문항 모수를 지정해줘야 하는 문제가 있어 베이지언 추정을 이용하였다. 베이지언 추정시 문항모수와 인지프로파일에 대한 사전분포(prior information) 설정은 두 가지 방안을 사용하였다. 첫 번째 방안은 모든 문항모수와 프로파일 분포에 단일분포(uniform distribution)를 가정하였고(BN-uniform), 두 번째 방안은 2학년 데이터에서 얻은 문항모수와 프로파일 분포를 Levy & Mislevy(2016)을 따라 사전정보로 설정하였다. 이후 모집단 데이터로부터 크기가 N=100, 150, 200인 샘플을 무작위로 100회 추출하여 세 가지 다른 추정법으로 문항 모수와 인지프로파일 모수를 추출한 후 모집단 데이터의 결과와 차이를 평균하여 다음과 같이 산정하였다.

$$MABg = Mean\,Absoulte\,Bias\,(guessing) = \frac{1}{K}\sum_{k=1}^{K}|\hat{g} - g| \tag{9}$$

$$MABs = Mean\,Absolute\,Bias\,(slip) = \frac{1}{K}\sum_{k=1}^{K}|\hat{s} - s| \tag{10}$$

$$PCV = \frac{1}{N}\sum_{i=1}^{N} I(\hat{a}_i - a_i) \tag{11}$$

위 식 (9), (10)에서 $\hat{g}$, $\hat{s}$ 는 N=100, 150, 200 크기의 데이터를 무작위 추출하여 세 가지 추정방법을 적용하였을 때 문항 모수 추정치이며 g, s는 모집단 추정치이다. 또한 식 (11)의 PCV(proportion of correct attribute vectors)는 위의 모의 실험 데이터에서 각 수험자의 프로파일이 모집단 데이터와 실험 데이터에서 일치하는 경우를 셈하여 그 비율을 결정하는 것이다.

4. 결과

모집단 데이터에 세 가지 다른 추정 방식으로 DINA 모형을 적용하여 산출한 문항 모수치는 '표 3'에 제시되어 있다. 모집단 분포에 대해 몇 가지 특징을 살펴볼 수 있는데, 첫째로 두 문항 모수에 대한 추정 방식간의 차이가 크지 않다. 대부분 차이가 소수점 둘째 자리에서 약간의 차이가 있음을 볼 수 있다. 둘째, 일부 문항의 추측 모수(guessing) 값이 상

당히 커서(예: q05, q06, q08, q09 등) 추측하여 정답을 맞추는 확률이 30%를 넘는 문항의 숫자가 적지 않음을 보여주고 있다. 셋째, 상당수의 문항에서 부주의 오류 모수(slip) (정답을 맞추기 위해 필요한 인지요인 을 다 갖추었음에도 불구하고 실수로 오답할 확률)가 매우 커서 문항의 텍스트, 질문, 답지의 언어적 특성이나 내용을 재검토할 필요가 있다. 넷째, 문항의 추측모수 범위보다 부주의 오류 모수의 범위가 일반적으로 협소한 것으로 알려져 있는데, 이 데이터에서는 그 반대 현상을 보여 그 원인에 대한 추가적인 조사가 필요하다.

표 3. 추정방법에 따른 모집단 문항모수

문항	CDM		BN-CDM(uniform)		BN-CDM(informative)	
	guessing	slip	guessing	slip	guessing	slip
q01	.127	.898	.117	.863	.124	.874
q02	.167	.728	.143	.638	.147	.642
q03	.163	.707	.150	.698	.162	.667
q04	.157	.487	.154	.497	.152	.512
q05	.361	.078	.340	.082	.321	.083
q06	.342	.117	.332	.109	.328	.109
q07	.207	.446	.145	.408	.157	.398
q08	.414	.121	.408	.116	.403	.115
q09	.389	.098	.301	.043	.270	.047
q10	.176	.709	.145	.618	.177	.610
q11	.218	.314	.150	.226	.179	.207
q12	.388	.073	.250	.031	.206	.036
q13	.052	.057	.133	.076	.131	.087
q14	.166	.530	.135	.433	.174	.409
q15	.169	.449	.128	.347	.166	.307
q16	.219	.252	.153	.135	.188	.104
q17	.114	.460	.163	.372	.193	.310
q18	.155	.414	.113	.340	.180	.279
q19	.024	.219	.045	.019	.004	.094
q20	.165	.747	.126	.743	.141	.746

다섯 개의 인지요인 조합으로 구성된 32개의 인지프로파일 상대도수 분포는 '표 4'에 제시되어 있다. '표 4'는 인지프로파일 분포의 전형적인 형태를 보여주는데 모수 추정 방식과 상관없이 가장 낮은 능력(00000 인지프로파일)과 가장 높은 능력(11111 프로파일)이 가

장 많은 부분을 차지하며 기타 다른 특성을 갖는 인지프로파일이 산발적인 분포를 보인다. 다섯 개의 인지요인 중 하나 이하의 인지요인만 완전학습을 보이는 경우를 BN-CDM (informative) 추정방법에서 찾아보면 36.6%를 차지하고 있어 상당한 보완 대책이 필요함을 보여주고 있다. 또한 다섯 개의 인지요인 중 한 개 이하에서만 불완전학습을 보이는 경우도 39.4%를 차지하고 있어 이 그룹의 학생을 위한 특별 프로그램도 필요하다.

'표 4'에서 완전학습 프로파일(11111)에 대한 BN-CDM(uniform) 추정 방법과 개별 인지요인 중 대의 파악(gist)에 대한 BN-CDM 추정 결과는 측정의 관점에서 주목할 만하다. 대부분의 다른 인지프로파일 추정에서 BN-CDM과 전통적인 CDM 추정 결과는 그 차이가 크지 않았는데, 완전학습 프로파일에 대해서는 BN-CDM(uniform)이 다른 두 추정방법도다 현저히 낮은 확률을 제시하고 있다. 물론 이 연구에서는 모집단 모수를 1학년 학생 전체 데이터에서 도출된 값으로 설정하였기 때문에 어느 정도의 오차가 있을 수 있지만 다른 추정치와의 차이가 통계적으로($p < .001$) 그리고 실질적으로 큰 차이가 있다. 이러한 차이에 대한 근본적 원인을 현재 알 수 없지만 베이지언 추정에서 단일분포를 사전정보로 설정하는 데서 유래할 가능성도 있을 수 있다. 또한 '대의 파악' 인지요인에 대한 BN-CDM의 추정치가 전통적인 CDM 결과와 현격한 차이를 보이는데, BN-CDM의 경우에는 사전정보 설정 차이에 상관없이 유사하나 전통적인 CDM과는 큰 차이가 있어 모수 추정상의 기술적인 문제를 포함한 좀 더 세밀한 분석과 해석이 필요하다.한편, '표 4'의 하단에는 개별 인지요인의 완전학습률을 보여주고 있는데 5개 인지요인 전체에 대해 완전학습률이 모든 모형에 걸쳐 높지 않다. 또한 모형마다 상대적 난이도에서 차이를 보이는데, 전통적 CDM의 경우 문법과 대의 파악 영역이 어려웠으며, BN-CDM의 경우 대의 파악이 가장 어려운 영역임을 보여주고 있다.

'표 3'과 '표 4'의 문항 모수와 인지프로파일 분포를 모집단으로 하여 소규모 샘플에서 문항모수 추정치와 인지프로파일 분류 일치도(PCV)를 비교한 결과는 '표 5'에 제시되어 있다. '표 5'는 소규모 샘플에서 세 가지 추정 방식에 대한 작지만 일관된 몇 가지 패턴을 보여주고 있다. 첫째, 문항 모수와 PCV 추정에 있어 최대우도추정(maximum likelihood estimation) 방법에 의존하는 전통적 CDM의 편차가 가장 크며, BN-CDM(informative)이 일관되게 가장 작은 편차를 보이고 BN-CDM(uniform)은 두 가지 추정방법의 중간값을 내놓고 있다. 그러나 BN-CDM(informative)의 이러한 특성은 사전정보 설정에서 현재 모집단 데이터보다 더 큰 1학년 시험 자료의 결과를 그대로 사전정보로 설정한 결과이므로 BN-CDM (informative)이 언제나 이런 특성을 보이는지는 좀 더 많은 조사가 필요하다. 둘째, 샘플 수가 증가함에 따라 전통적인 CDM은 편차가 지속적으로 감소하는 반면, BN-CDM의 샘플 수 증가에 따른 편차 감소폭은 크지 않다. 이는 소규모 샘플에 기반한 BN-CDM의 베이지언 추정에서 사전정보의 영향력이 데이터의 정보를 압도하는 것으로 보여진다. 셋째, 문항 모수 중에서는 추측도 모수의 편차가 부주의 오류모수 편차보다 추정 방식에 상관없이 언제나 크

다. 이는 CDM 문항 모수 추정에서 일반적인 현상으로(Sorrel et al., 2023) '표 3'에서 볼 수 있는 부주의 오류모수의 큰 변량이 예외적인 현상임을 보여준다.

표 4. 모집단 인지프로파일 분포

인지프로파일	CDM	BN-CDM(uniform)	BN-CDM(informative)
00000	**.111**	**.177**	**.183**
10000	.006	.011	.017
01000	.002	.019	.012
00100	.073	.054	.017
00010	.051	.044	.035
00001	**.096**	**.077**	**.102**
11000	.012	.005	.014
10100	.007	.002	.000
10010	.000	.016	.016
10001	.000	.000	.000
01100	.000	.000	.000
01010	.012	.033	.025
01001	.051	.040	.044
00110	.000	.000	.000
00101	.012	.004	.012
00011	.059	.059	.059
11100	.002	.012	.002
11010	.007	.069	.006
11001	.015	.012	.015
10110	.000	.000	.001
10101	.000	.000	.000
10011	.008	.012	.007
01110	.004	.000	.004
01101	.000	.000	.000
01011	.041	.037	.042
00111	.030	.000	.030
11110	.038	.047	038
11101	.004	.005	.004
11011	.019	**.065**	.019

인지프로파일	CDM	BN-CDM(uniform)	BN-CDM(informative)
10111	.000	.000	.000
01111	.017	.000	.017
11111	**.315**	**.194**	**.316**
grammar	.436	.453	.477
vocabulary	.543	.543	.547
gist	.507	**.321**	**.239**
specifics	.605	.579	.588
inference	.670	.508	.584

이 연구의 주요 주제인 소규모 샘플에서 CDM 적용시 발생할 수 있는 편차와 정확도에 관한 내용은 '표 5'에 제시되어 있다. 우선 '표 5'의 평균 표준 편차가 실제 CDM 적용에서 어떤 의미를 갖는지 살펴볼 필요가 있다. '표 5'의 전통적인 CDM 추정 방법에 따르면 문항의 추측도 모수의 경우, 편차값이 샘플 수에 따라 .061~ .085의 분포를 보인다. 이런 정도의 편차는 '표 3'의 추정방법에 차이에 따른 문항 모수 추정값에도 흔히 볼 수 있으나, '표 5'의 값은 '표 3'과 같은 추정 과정을 100회 반복하여 얻은 평균값임에 주의해야 한다. 양적 연구에서 모수 추정의 편차값은 정규분포를 일반적으로 가정하므로 표준오차를 .02라고 가정하면(실제 문항 모수의 표준오차는 .008 ~ .027임) 95% 신뢰구간이 .100 ~.124이므로 상당한 편차가 있으며, 이러한 편차가 학생의 인지프로파일 분류의 신뢰도에 미치는 영향이 어느 정도인지는 좀 더 정확한 연구가 필요하다. 이와 대조적으로 BN-CDM(informative)의 편차는 CDM의 약 1/4 정도로 베이지언 추정법의 효과를 확인할 수 있지만, 이는 모집단 모수에 근접하는 사전정보를 설정했기 때문이며 실제 이러한 정보가 없는 경우 기대하기 어려운 수치일 가능성이 있다.

표 5. 샘플 크기에 따른 문항 모수와 인지 프로파일 정확도 편차 비교

	guessing			slip			PCV		
	N=100	N=150	N=200	N=100	N=150	N=200	N=100	N=150	N=200
CDM	.085	.078	.071	.061	.060	.054	.653	.675	.690
BN-CDM(uniform)	.038	.034	.031	.027	.023	.020	.712	.731	.737
BN-CDM(informative)	.024	.022	.022	.021	.021	.018	.796	.796	.801

'표 5'의 PCV는 앞의 추측도, 부주의 오류 모수 편차를 나타내는 것과 대조적으로 정확도를 나타내므로 주의할 필요가 있다. 이 수치는 모집단 데이터의 인지프로파일 분류와 소규모 샘플의 인지프로파일 분류 정확도 정도를 나타내는 것으로 일종의 신뢰도로 해석할

수 있다. 전통적인 CDM의 경우 분류 일치도가 .653 ~.690이므로 높지 않으나 소규모 샘플과 소규모 시험(20문항)의 정확도로 수용할 정도이며, BN-CDM(informative)은 약 .80 정도의 분류 일치도를 보여 상당한 정확성을 갖고 있음을 알 수 있다. 그러나 모집단 데이터의 분류 일치도가 100%가 아님을 감안하면, 과대 해석의 위험이 있다. 만일 모집단 분류 일치도를 90%로 가정하고, 소규모 샘플의 경우 전통적인 CDM과 BN-CDM(informative)의 분류 일치도를 각각 .7, .8로 가정하면 실제 분류 일치도는 .63 ~.72이므로 많은 개선의 여지가 남아있으며, 특히 BN-CDM(informative)의 경우 사전정보가 과도하게 정밀하게 설정된 결과일 가능성이 많아 좀 더 다양한 사전정보 설정에 따른 민감도 검정이 요구된다.

마지막으로 최근 CDM 관련 문헌에서는 인지요인간 과도한 상관관계로 인한 CDM 적용 결과의 타당성에 많은 의문을 제기하고 있다(예: Sessom & Henson, 2018). 이 연구의 '표 4'에서 볼 수 있는 것처럼 많은 인지프로파일이 모집단 자료에서도 존재하지 않을 수 있는데, 실제 존재하지 않거나 분간하기 어려운 인지요인에 대한 CDM 적용을 예방하기 위한 여러 가지 가능성이 시험 데이터의 다차원성 검사(dimensionality testing) 위주로 논의되고 있다(Najera et al., 2000). 이 연구에서는 5-요인 인지모형의 타당성에 대한 측정학적 근거를 위해 Sorrel(2021)이 제안한 병렬분석(parallel analysis)을 실시 하였고 그 결과는 '그림 4'에 제시되어 있다. 분석에 따르면 주성분 분석의 경우 3개의 요인을 추출할 수 있으며, 요인분석의 경우 5개의 유의미한 요인을 상정할 수 있어 이 연구의 5-요인 읽기 모형이 측정학적 타당성을 갖추고 있음을 보여준다.

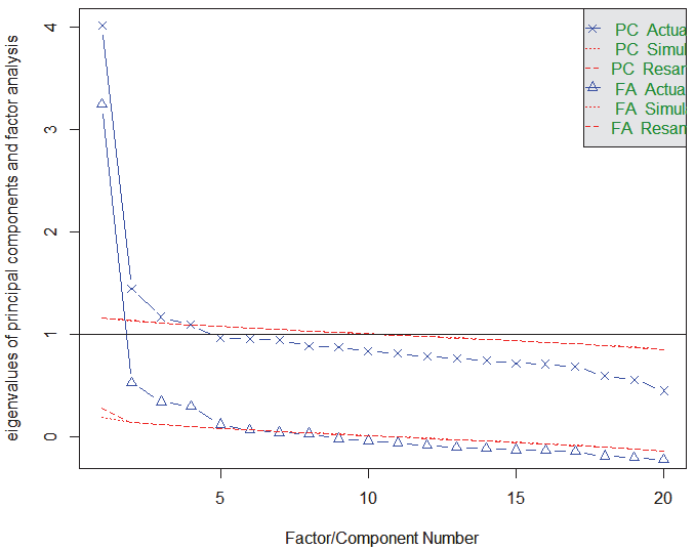

그림 4. 병렬분석(parallel analysis)

5. 논의 및 결론

전통적인 총합평가에 대한 대안으로써 CDM이 도입된지 한 세대가 지났으며 CDM의 측정학적 특성에 대한 많은 연구 결과가 축적되었다. 그러나 지금까지 CDM은 주로 대규모 시험 데이터에 기초하여 측정학적 특성 탐구에 초점을 두고 있어 소규모 샘플과 시험을 반복 시행하는 학교의 교수· 학습 상황과 많은 괴리가 있다. 최근 CDM 분야에서는 학교의 교수· 학습 상황에 맞는 CDM 모형 개발에 관심이 점증하고 있으며, 연구의 초점은 소규모 샘플에서 소규모 시험을 통해 CDM 본래의 목적을 이루기 위한 모형 선택과 시행 방법에 초점을 두고 있다. 이 연구에서는 근래 데이터 과학 분야에서 널리 쓰이는 베이지언 네트워크를 이용하여 소규모 샘플에 적용하여 봄으로써 CDM을 학교의 교수· 학습 상황에 적용할 수 있는지 조사하였다. 연구 결과 BN-CDM은 소규모 샘플 환경아래서 CDM을 적용할 수 있는 많은 가능성을 보여주었다. 문항 모수의 비교 결과를 보면 모든 조건에서(N=100, 150, 200) BN-CDM의 편차가 일률적으로 더 적으며, 인지프로파일 분류 일치도에서도 BN-CDM이 더 나은 결과를 보여주었다. 이러한 결과는 추정 과정을 국소적으로 시행하는 BN-CDM의 특성과 사전정보의 이용을 체계적으로 시행할 수 있는 베이지언 추정 방법의 특성에 기인한다고 볼 수 있다.

이 연구에서 살펴본 읽기 시험 데이터에 대한 BN-CDM 적용은 다음 몇 가지 사항을 고려할 필요성을 제기한다. 첫째, 영어 읽기의 인지적 과정, 특히 구성요소와 요소간 관계 또는 위계에 대한 장기적인 연구가 필요하다. BN-CDM의 적용과 관련하여 모형 개발이 문제가 되는 이유는 인지요인의 위계구조에 아무런 관계 설정을 하지 않는 전통적인 CDM과 달리 BN-CDM은 어떤 형태로든지 모형을 설정해야 하는 특성이 있기 때문이다. BN-CDM의 장점인 국소지향적 추정은 선택된 모형의 그래프 인수분해에 크게 의존하기 때문에 명확한 읽기 모형이 설정이 전제되어야 한다. 그러나 현재 읽기 영역의 모형 개발은 주로 L1 읽기에서 진행되고 있고 L2 영역에서는 관련 인지요인을 상정하고 있으나 이 요인들의 관계를 입증하기 위해서는 상당한 연구 결과 축적이 필요한 실정이다(Grabe, 2009). 이 연구에서는 선행연구 검토를 통하여 영어 읽기 영역 CDM에서 가장 일반적으로 적용되는 요인들로 모형을 구성하였으나 다른 모형의 개발과 비교 과정이 요구된다. 예를 들면, 영어 읽기 영역에 관한 많은 연구에서 문법과 어휘 능력이 학생의 영어 능력이 향상됨에 따라 통합된다고 주장하고 있다(Alderson et al., 2014; Harding et al., 2015). 이러한 주장은 문법과 어휘 능력에 관한 임계 모형, 즉 두 가지 능력이 어느 기준 정도를 넘어서면 이후에는 다른 능력이 주로 영향을 미친다는 것으로 BN-CDM은 이러한 모형을 용이하게 적용할 수 있어 읽기 모형의 실증적 검증에도 기여할 수 있을 것이다. 한편 이 연구에서 고려한 인지요소의 구성과 위계는 매우 제한적이어서 단지 5가지 요인을 고려했을 뿐이나, 관련 문헌에서 쉽게 찾아볼 수 있는 동기(Grabe, 2009), L1 읽기 능력(Olsen et al., 2016), 배경 지식(Lee, 1986),

일반적 추리 능력(Droop & Verhoeven, 2003) 등을 포함한 실제 영어학습자의 읽기 능력을 어느 정도 포괄할 수 있는 모형의 개발과 실제 적용을 통한 검증이 필요하다.

둘째, 이 연구와 관련하여 CDM 또는 BN-CDM 적용을 위한 또 다른 과제는 소규모 샘플에서 최적 모형 선택 문제이다. 지금까지 영어 읽기 분야의 CDM 관련 문헌에서 인지진단모형 선택을 보면 분석 프로그램의 디폴트 모형을 선택하거나(예: RUM(NC-RUM, C-RUM)관련 모형 선택), 포화모형(예: GDM, LCDM, GDINA)을 최적 모형으로 선택하는 경우를 흔히 볼 수 있다(Li et al., 2015; Ravand & Robitzsch, 2018). 그러나 이 연구들은 대부분 대규모 데이터에 기초하고 있으며, 포화모형의 경우 모수 숫자뿐만 아니라 인지요인 간 교차작용(interaction effects)에 따른 모수 추정과 해석 문제가 발생할 수 있다(Yi, 2018). 소규모 샘플에서 CDM의 모형 축소는 불가피한 것이며 문제는 이러한 모형의 축소가 인지프로파일 분포 추정에 어느 정도 영향을 미치는지 알려져 있지 않다는 것이다. 이 연구에서는 BN-CDM의 적용에서 모형의 다양성에 따른 추정 결과를 비교하지 않고, 교실 상황에서 CDM 적용 모형 선택연구에서 빈번하게 채택되는 DINA 모형에 국한하여다. 그러나 적정 인지요인 수의 결정과 각 모형의 선택에 따른 인지프로파일의 편차와 신뢰도 변화 정도는 CDM의 기반 조성에 반드시 선결되어야 할 과제이다.

셋째, CDM 적용과 관련하여 중요한 하나의 문제는 진단평가용 문항 개발에 관한 것이다. 지금까지 영어 평가의 CDM 연구와 본 연구에서도 기존에 이미 개발된 시험 문항을 이용하였는데 교실의 교수·학습 상황에서는 교사가 문항을 개발해야 하는 과제가 남아 있다. 이 연구에서 볼 수 있듯이 대학의 어학교육원에서 다수의 평가전문가가 공동 개발한 시험의 경우에도 추측도 모수나 부주의 오류모수가 매우 큰 경우가 적지 않음을 알 수 있다. 학교의 교수·학습 환경에서 교사 개인이 어떻게 다수의 문항을 개발하며 개발된 문항의 질적 관리를 유지할 수 있는지는 매우 어려운 과제이다. 그 이유는 영어 평가의 각 영역에서 다양한 난이도에 걸쳐 변별도가 높은 문항을 지속적으로 개발해야 하기 때문이다. CDM 적용을 위해서는 모형 선택이나 분석과 같은 측정학적 문제뿐만 아니라 문항 개발에 관한 지속적인 자원의 투입과 관심이 필요한 이유이기도 하다. 이와 관련하여 우리의 교육 상황은 공통의 국가 교육과정을 이수하며, 특히 지역사회에서는 동질적인 교육환경이 조성되어 있으므로 교사 개인보다는 지역의 교사가 연합하여 문제은행을 개발 공유하고 분석 결과를 상호 간 선행 정보로 이용하여 분석의 정밀성을 더해가는 체계를 고려해 볼 수 있다. 특히 소규모 샘플을 이용한 CDM의 안정적 시행을 위해서는 반복되는 시험과 분석 자료의 누적으로부터 발생하는 정보(soft evidence)를 최대한 활용해야 할 필요성이 있는데, BN-CDM은 태생적으로 이러한 기능을 갖추고 있어 문제은행의 공동 개발과 분석 결과 공유는 교사 개인이 인지진단 평가 시스템을 운영해야 하는 버거운 과정으로부터 독립할 기회를 제공할 수 있을 것이다.

마지막으로 이 연구의 큰 제한점이며 동시에 대부분의 CDM 적용 연구의 문제점은 진

단시험 결과에 대한 피드백의 부제이다. 이 연구 또한 소규모 샘플에서 BN-CDM의 측정학적 특성에 초점을 두어 인지프로파일에 대한 피드백을 고려하지 않았으나, CDM 시행의 궁극적인 목적 실현을 위해서는 이 부분이 반드시 충족되어야 한다. 피드백 제공과 관련하여 다음 몇 가지 사항을 고려할 필요가 있다. 첫째, 의미 있는 피드백 제공을 실현하기 위해서는 이 연구에서 고려한 것처럼 모형의 크기가 작아야 한다. 이 연구에서는 5개의 인지요인을 상정하였으므로 이론적으로 32개의 인지프로파일이 가능하며 실제 상당한 빈도수를 갖는 프로파일은 불과 3~4개 정도이다. 그러므로 피드백은 3~4개의 인지프로파일에 집중하되, CDM의 기본적 전제인 개인의 학습을 위한 진단정보 제공에 충실하기 위해서는 빈도수가 낮은 경우라도 이에 대한 세밀한 피드백 제공을 고려해야 할 것이다. 둘째, 피드백 제공에서 즉시성 문제와 빈도수가 낮은 인지프로파일에 대한 피드백 제공의 효율성 문제를 고려할 때, BN-CDM을 온라인에서 실행하는 문제를 적극적으로 고려할 필요가 있다. 현재 우리나라 학교의 교수·학습 상황에서 개별 교사에 의한 진단평가의 피드백 제공은 현실적으로 많은 어려움이 있으므로 시행과 분석, 피드백 제공 등을 일괄적으로 시행할 수 있도록 BN-CDM을 컴퓨터 적응검사 형식으로 개발할 필요가 있으며 이러한 과제 수행을 위해서는 다시 지역사회 교사의 공동연구와 협업을 통한 인지진단 체계의 수립이 보다 더 현실적인 대안으로 판단된다.

참고문헌

Afflerbach, P., Cho, B. Y., Kim, J. Y., Crassas, M. E., & Doyle, B. (2013). Reading: What else matters besides strategies and skills?. *The Reading Teacher, 66*(6), 440-448.

Akbay, L., & de la Torre, J. D. L. (2020). Estimation approaches in cognitive diagnosis modeling when attributes are hierarchically structured. *Psicothema*, *32*(1), 122-129

Alderson J. C. (1990a). Testing reading comprehension skills (Part Two). *Reading in a Foreign Language, 7*, 465–503.

Alderson J. C. (1990b). Testing reading comprehension skills (Part One). *Reading in a Foreign Language, 6*, 425–438.

Alderson J. C. (2000). Assessing reading. Cambridge: Cambridge University Press.

Alderson, J. C. (2005). Assessing reading. Cambridge, MA: Cambridge University Press.

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5*, 253-70.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). Bayesian networks in educational assessment. Springer.

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25*, 133-150.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*, 395-414.

Bradshaw, C. P., Milam, A. J., Furr-Holden, C. D. M., & Johnson, L. S. (2015). The School Assessment for Environmental Typology (SAfETy): An observational measure of the school environment. *American Journal of Community Psychology, 56*, 280-292.

Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., & Phelps, M. (1998). Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I verbal: Sentence completion section (ETS Research Report No. RR-98-23). Princeton, NJ: Educational Testing Service.

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), 119–157.

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*(2), 84-95.

Chen, J., & de la Torre, J. (2012). An extension of the G-DINA model for polytomous attributes. Paper presented at the annual meeting of American Educational Research Association, Vancouver.

Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences, 10*(22), 8196.

Choi, Y., & Zhang, D. (2021). The relative role of vocabulary and grammatical knowledge in L2 reading comprehension: A systematic review of literature. *International Review of Applied Linguistics in Language Teaching*, *59*, 1-30.

Cowell, R. G., Dawid, P., Lauritzen, S. L. & Spiegelhalter, D. J. (2007). Probabilistic networks and expert systems: Exact computational methods for Bayesian networks. Springer Science & Business Media.

Culbertson, M. J. (2016). Bayesian networks in educational assessment: The state of the field. *Applied Psychological Measurement, 40*(1), 3-21.

Davey, B. (1988). Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *The Journal of Experimental Education, 56*(2), 67-76.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-Matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447–468.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115-130.

DiBello, L. V., Stout W. F., & Roussos L. A. (1995). Unified cognitive/ psychometric diagnostic assessment likelihood-based classification techniques. In Nichols P. D., Chipman S. F., Brennan R. L. (Eds.), Cognitively diagnostic assessment (pp. 361-389). Routledge.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods, 50*(1), 195-212.

Finkelman, M., Kim, W., Weissman, A., & Cook, R. (2014). Cognitive diagnostic models and computerized adaptive testing: Two new item-selection methods that incorporate response times. *Journal of Computerized Adaptive Testing, 2*(4), 59-76.

Gottardo, A., & Mueller, J. (2009). Are first- and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology, 101*, 330–344.
doi: 10.1037/a0014320

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6-10.

Grabe, W. (2009) Reading in a second language: moving from theory to practice. Cambridge University Press, Cambridge.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing, 32*(3), 317-336.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2*(2), 127–160.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 23*, 403–430.

Hughes, A. (2003) Testing for Language Teachers. 2nd Edition, Arthur Hughes, Cambridge.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion Model to LanguEdge assessment. *Language Testing, 26*(1), 31–73.

Jang, E. E., Dunlop, M., Wagner, M., Kim, Y. H., & Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. *Language Learning, 63*(3), 400-436.

Kasai, M. (1997). Application of the rule space model to the reading comprehension section of the test of English as a foreign language (TOEFL). University of Illinois at Urbana-Champaign.

Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing, 28*(4), 509–541.

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227-258.

Koller, D., & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), Special language: From humans to thinking machines (pp. 316–323). Clevedon, UK: England: Multilingual Matters.

Lee, J. F. (1986). Background knowledge and L2 reading. *The Modern Language Journal, 70*(4), 350-354.

Lee, J. W. (2016). The role of vocabulary and grammar in different L2 reading comprehension measures. *English Teaching, 71*, 79–97.

Lee, B. Y., & Shin, S. K. (2020). Doable and practical: A validation study of classroom diagnostic tests. *Journal of Asia TEFL, 17*(2), 363.

Lee, Y-W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*(3), 239-263.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). Cognitive diagnostic assessment for education: Theory and applications. Cambridge University Press.

Levy, R., & Mislevy, R. J. (2017). Bayesian psychometric modeling. CRC Press.

Li, Y., Huang, C., & Liu, J. (2023). Diagnosing primary students' reading progression: Is cognitive diagnostic computerized adaptive testing the way forward? *Journal of Educational and Behavioral Statistics, 48*(6), 842-865.

Li, Y., Zhen, M., & Liu, J. (2021). Validating a reading assessment within the

cognitive diagnostic assessment framework: Q-matrix construction and model comparisons for different primary grades. *Frontiers in Psychology,* 2021 Dec 16; 12, 786612.

Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, 33*, 391–409.

Lin, Q., Xing, K., & Park, Y. S. (2020) Measuring skill growth and evaluating change: unconditional and conditional approaches to latent growth cognitive diagnostic models. *Frontiers in Psychology, 11*, 2205. doi: 10.3389/fpsyg.2020.02205

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow, 9*, 17-46.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing, 10*(3), 211-234.

Ma, W., & Jiang, Z. (2021). Estimating cognitive diagnosis models in small samples: Bayes modal estimation and monotonic constraints. *Applied Psychological Measurement, 45*(2), 95-111.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling, 23*(5), 750–773.

Morvay, G. (2012). The relationship between syntactic knowledge and reading comprehension in EFL learners. *Studies in Second Language Learning and Teaching, 2*(3), 415-438.

Nájera, P., Abad, F. J., Chiu, C-Y., & Sorrel, M. A. (2023). The restricted DINA model: A comprehensive cognitive diagnostic model for classroom-level assessments. *Journal of Educational and Behavioral Statistics, 48*(6), 719-749.

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59-82.

Neapolitan, R. E. (2004) Learning Bayesian Networks. Prentice Hall, Upper Saddle River, NJ.

Brevik, L. M., Olsen, R. V., & Hellekjær, G. O. (2016). The complexity of second language reading: Investigating the L1-L2 relationship. *Reading in a Foreign Language, 28*(2), 161-182.

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: recent developments, practical issues, and prospects. *International Journal of Testing, 20*,

24–56. doi: 10.1080/15305058.2019.1588278

Pan, Y., & Zhan, P. (2020) The Impact of sample attrition on longitudinal learning diagnosis: A Prolog. *Frontiers in Psychology, 11*, 1051. doi: 10.3389/fpsyg.2020.01051

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782–799.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). Diagnostic measurement: Theory, methods, and applications. New York, NY: Guilford Press.

Sawaki, Y., Kim, H-J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*(3), 190-209.

Sen, S., & Cohen, A. S. (2021). Sample size requirements for applying diagnostic classification models. *Frontiers in Psychology, 11*, 621251. doi: 10.3389/fpsyg.2020.621251

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: a literature review and critical commentary. *Measurement Interdisciplinary Research and Perspectives, 16*, 1-17.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24*(1), 99-128.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement, 67*(2), 239-257.

Sorrel, M. A., Abad, F. J., & Nájera, P. (2021). Improving accuracy and usage by correctly selecting: The effects of model selection in cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 45*(2), 112-129.

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement, 4*1(8), 614–631.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*, 32–71.

Sun, X., Gao, Y., Xin, T., & Song, N. (2021) Binary restrictive threshold method for item exposure control in cognitive diagnostic computerized adaptive testing. *Frontiers in Psychology, 12*, 517155. doi: 10.3389/fpsyg.2021.517155

Susoy, Z., & Tanyer, S. (2019). The role of vocabulary vs. syntactic knowledge in L2 reading comprehension. *Eurasian Journal of Applied Linguistics, 5*, 113-130. doi: 10.32601/ejal.543787

Yi, Y. -S. (2017). In search of optimal cognitive diagnostic model(s) for ESL grammar test data. *Applied Measurement in Education, 3*0(2), 82-101.

Van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Snellings, P., Simis, A., et al. (2003). Roles of linguistic knowledge, metacognitive knowledge and processing speed in L3, L2 and L1 reading comprehension: A structural equation modeling approach. *International Journal of Bilingualism, 7*, 7–25.

Wang, N., & Almond, R. (2019). Bayesian model checking in cognitive diagnostic models. *Behaviormetrika, 46*(2), 371-388.

Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British …. British Council/IDP Australia Research Reports. 9.

Zhan, P., Jiao, H., Man, K., & Wang, L. (2019). Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics, 44*(4), 473-503.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *Modern Language Journal, 96*, 558–575.

최세일
전남대학교 사범대학 영어교육과 강사
61452 광주광역시 북구 용봉로 77
전화: (062)431-8788
이메일: csieagles@gmail.com

Received on November 30, 2023
Revised version received on December 21, 2023
Accepted on December 31, 2023