과학교육연구지, 47권 2호, pp. 211-224 (2023.8)
ISSN : 1225-3944
Jour. Sci. Edu, Vol. 47, No, 2, 211-224 (2023.8)
eISSN : 2733-4074

다변량일반화가능도 이론을 적용한 자동문항생성 기반 평가에서의 신뢰도 탐색

정진민 · 김성연[1*]
아이오와대학교 · [1]인천대학교

Exploring the Reliability of an Assessment based on Automatic Item Generation Using the Multivariate Generalizability Theory

Jinmin Chung · Sungyeun Kim[1*]
University of Iowa · [1]Incheon National University

Abstract : The purpose of this study is to suggest how to investigate the reliability of the assessment, which consists of items generated by automatic item generation using empirical example data. To achieve this, we analyzed the illustrative assessment data by applying the multivariate generalizability theory, which can reflect the design of responding to different items for each student and multiple error sources in the assessment score. The result of the G-study showed that, in most designs, the student effect corresponding to the true score of the classical test theory was relatively large after residual effects. In addition, in the design where the content domain was fixed, the ranking of students did not change depending on the item types or items. Similarly, in the design where the item format was fixed, the difficulty showed little variation depending on the content domains. The result of the D-study indicated that the original assessment data achieved a sufficient level of reliability. It was also found that higher reliability than the original assessment data could be obtained by reducing the number of items in the content domains of operation, geometry, and probability and statistics, or by assigning higher weights to the domains of letters and formulas, and function. The efficient measurement conditions presented in this study are limited to the illustrative assessment data. However, the method applied in this study can be utilized to determine the reliability and to find efficient measurement conditions for the various assessment situations using automatic item generation based on measurement traits.

keywords : assessment, automated item generation, multivariate generalizability theory, reliability

Ⅰ. 서론

최근 인공지능, 빅데이터, 클라우드, 사물인터넷, 가상융합기술, 블록체인 등과 같은 4차 산업혁명의 주요 기술의 급속한 발전으로 산업, 사회, 경제 및 문화는 디지털 기반으로 빠르게 변화하고 있다(Lee, 2022). 이러한 변화는 교육 분야에서도 예외는 아니며 전 세계적으로 학습 효과를 높이기 위해 디지털 기술을 적극적으로 활용하고 있다. 영국은 영국교육기자재협회(BESA) 지원을 통해 학교를 위한 에듀테크 플랫폼인 'LendED'를 구축하고, 독일의 연방교육부는 '디지털 팍트(DigitalPakt Schule)' 사업을 추진하고, '디지털 교육 이니셔티브'를 발표하였다. 또한 미국의 연방 정부는 '국가교육기술계획(National Education Technology)'을 발표하여 미국 교육 내 기술 활용 방향을 제시하고 있고, 일본의 문부과학성은 '기가(GIGA) 스쿨 구상' 정책을 통해 '학습e포털'을 구축하였으며, 에스토니아는 '타이거 리프 프로젝트'를 시작으로 디지털 교육환경을 선제적으로 조성하였으며, 싱가포르는 디지털 기술 적용 및 활용을 목표로 하는 'EdTech Plan'을 발표하였다(MOE, 2023). 우리나라의 교육부에서도 모든 학생을 인재로 키우기 위한 맞춤 교육 실현을 위해 AI 등 첨단 기술을 활용하여 교육의 질을 제고하고, 디지털 대전환에 따른 공교육에서의 변화 노력

* 교신저자: 김성연 (syk@inu.ac.kr)
** 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2019R1F1A1059437, No.2022R1A2C1010310).
*** 2023년 6월 26일 접수, 2023년 8월 5일 수정원고 접수, 2023년 8월 5일 채택
http://dx.doi.org/10.21796/jse.2023.47.2.211

과 시도를 주요 배경으로 하는 '디지털 기반 교육혁신 방안(MOE, 2023)' 및 '코로나 이후, 미래교육 10대 정책과제 시안'을 발표하였다. 구체적으로 1)미래형 교육과정 개편, 2)새로운 교원제도 마련, 3)학생 중심 미래형 학교 조성, 4)성장 지원 교육안전망 구축, 5)협업공유를 통한 대학지역의 성장, 6)미래사회 핵심 인재양성, 7)고등 직업교육 내실화, 8)전 국민, 전 생애 학습권 보장, 9)디지털 전환 교육 기반 마련, 그리고 10)협력적 교육 거버넌스 구축이 제시되었다. 이 중 모든 학생의 기초학력 책임을 보장하기 위한 교육안전망 구축 정책과제에서는 학교교육을 통한 기초학력 지원 강화를 위해 원격수업·자율학습이 가능한 온라인 콘텐츠에 자동문항생성(Automatic item generation) 기능을 추가하여 유사한 문항의 학습을 지원할 수 있도록 하였다(MOE, 2020).

자동문항생성이란 기계가 기호 조작할 수 있는 형태로 문항모형(Item model)을 제작하여 컴퓨터가 자동으로 문항모형에 포함된 요소들을 정해진 규칙에 따라 생성함으로써 개별적인 문항이 아니라 문항모형을 통해 다양한 유형의 서로 다른 문항들을 생성하게 하는 것을 말한다. 문항모형이란 지식을 이루는 핵심 요소의 구조와 요소를 분석하여 모형화한 것이다(Gierl & Lai, 2013; Kim, 2023). 이처럼 문항이 자동문항생성과 같은 디지털 기술로 전환되면 전통적인 평가가 갖는 한계점인 제작 과정에 있어 몇몇 전문가의 개별적인 수작업 및 평가지를 읽고 응답하는 과정으로 이루어지는 비과학적이고, 고비용이며, 느리며, 시행에 있어 시험 보안에 취약하고, 연속적으로 시행하기 어려우며, 전반적인 평가 시스템을 학습 및 교육에 최적화하기 어려우며, 지능화하는 데 있어서의 어려움을 극복하는 데 활용될 수 있다(Choi, 2020). 이처럼 자동문항생성은 문항을 생성하는 데 필요한 시간과 자원을 줄일 수 있으며, 학생들이 특정 문항을 공유하거나 시험에서의 부정행위를 할 수 있는 가능성을 줄여주며, 다양하고 복잡한 평가 문항을 생성함으로써 학생들의 지식과 기술을 보다 정확하게 측정하여 시험 타당성을 확보할 수 있게 한다. 또한 학생 개인의 능력과 필요를 고려한 맞춤형 평가를 시행하는 것이 가능하므로, 학생들의 능력을 보다 정확하게 평가하여 평가의 효율성, 타당성, 보안성, 맞춤성 등을 향상시키는데 기여할 수 있다(Haladyna & Rodriguez, 2013).

특히 단계별 학습 및 문제은행 확보가 중요한 수학교육 분야에서 자동문항생성은 높은 관심을 보이고 있는데(Kang & Choi, 2020), 이는 문항 유형과 난이도를 조절하여 다양한 문항을 생성함으로써 학생들에게 여러 가지 수학적 개념과 능력을 향상시키는데 활용할 수 있기 때문이다. 또한 수학적 사고 과정과 논리적 추론을 반영한 문항을 생성하여 학생들이 이를 풀고 해결함으로써 수학적 사고 능력을 발전시킬 수 있으며, 수학 개념을 다양한 맥락과 문제 상황에 적용하는 문항으로 생성함으로써 수학 개념을 실제 상황에 적용하고 문제를 해결하는 데 필요한 응용력을 기르는데 활용할 수 있다. 또한 특정 주제나 영역에 관련된 문항을 자동생성함으로써 학생들에게 이와 관련하여 특화된 수학적 이해와 능력을 향상시킬 수 있으며, 자동문항생성을 통해 생성된 문항을 푸는 과정과 결과를 분석하여 학생들의 개별적인 성취도를 평가하고, 이를 바탕으로 학생들에게 필요한 개념과 문제해결 능력을 개선할 수 있는 개별화된 피드백을 제공할 수 있다.

이처럼 자동문항생성을 활용하여 평가를 수행할 때 학생들의 능력을 더 정확하게 추정하고, 이를 바탕으로 개별적인 피드백을 제공하기 위해서는 평가도구의 신뢰도가 확보되어야 한다. 그러나 자동문항생성을 적용한 연구들에서는 생성된 문항들에 대한 신뢰도 산출보다는 자동문항생성 기반의 시스템 개발(Choi, Kim, & Pak, 2018, Choi *et al.*, 2022; Kang & Choi, 2020; Lim, 2017), 자동문항생성 기반의 시스템 활용 예시(Kim, 2022, 2023; Oh, 2022), 생성된 문항들이 포함되는 학습 콘텐츠를 교육 현장에 적용하여 콘텐츠 전과 후의 학업성취도 결과를 분석(Jeong *et al.*, 2009), 문항의 내용 및 특성을 고려하여 생성된 문항들이 난이도 및 변별도에 차이가 있는지를 분석하는(Lim, 2017) 연구들이 주를 이루고 있다. 최근 Falcão *et al.* (2023)은 자동문항생성으로 생성한 검사 문항들의 질, 유용성 및 타당도를 검토하면서 문항반응이론을 바탕으로 문항정보함수와 검사정보함수를 통해 신뢰도를 산출하였다. 즉, 자동문항생성을 활용한 연구에서 신뢰도에 관한 연구는 많이 이루어지지 않았으며(Lim, 2017; Jeong, 2009), 측정오차의 다양한 요인을 반영하지 못하는 문항정보함수와 검사정보함수를 통해 신뢰도를 산출하고 있는 실정이다(Falcão *et al.*, 2023). 따라서 본 연구에서는 측정 상황에서 발생하는 다양한 오차 요인을 동시에 분석할 수 있는 일반화가능도이론을 활용하여 신뢰도를 산출하고자 한다.

일반화가능도이론은 단일오차요인만을 고려하는 기존 검사이론의 한계를 보완하며, 연구자가 상정한 측정 상황에서 발생할 수 있는 다양한 오차요인들을 동시에 분석함으로써 평가 점수에 기여하는 다양한 오차요인의 상대적인 영향력을 파악하고, 이를 바탕으로 적정 수준의 신뢰도에 도달할 수 있는 효율적인 측정 조건에 대한 정보를 제공한다. 구체적으로 일반화가능도이론은 G-연구(G-study)와 D-연구(D-study)로 구분된다. G-연구는 표집단위가 측정 대상인 모집단과

는 다르게 측정 대상의 측정 조건들에 대한 일반화과정을 포함하는 허용가능한 관찰 전집(universe of admissible of observation)과 관련된 조건들의 분산성분 추정값을 추정함으로써 오차요인들의 상대적인 영향력을 파악할 수 있게 하며, D-연구 설계의 조건을 결정하는 근거로 활용된다. D-연구는 G-연구의 분석 결과를 토대로 측정 대상과 일반화 전집(universe of generalization)의 정의에 따라 전집점수 분산, 오차점수 분산, 규준참조평가에서 활용되는 신뢰도인 일반화가능도계수 및 준거참조평가에서 활용되는 신뢰도인 의존도계수의 정보를 제공함으로써 효율적인 측정 조건을 결정할 수 있는 정보를 제공한다. 본 연구에서 활용하는 다변량일반화가능도이론은 일변량일반화가능도이론의 특수한 형태로 일변량일반화가능도분석을 시행하는 경우보다 고정국면과 임의국면이 혼합된 불균형 설계처럼 복잡한 경우에 평가 결과를 동시에 분석할 수 있어 상대적으로 쉽게 적용할 수 있다. 여기서 국면(facet)이란 측정 조건들의 집합으로 분산분석 설계에서의 효과 또는 요인과 유사한 개념이며, 각 측정 조건이 동일한 표본을 사용했는지 여부에 따라 고정국면과 임의국면으로 구분한다. 또한 일변량일반화가능도이론에서 제공하지 못하는 분산, 공분산 성분 추정치를 제공해주며, 이를 바탕으로 고정효과 국면의 수준별 전집점수에 가중치를 준 합성점수에 대한 오차점수 분산 및 신뢰도를 제공한다(Brennan, 2001a; Kim, 2001; Kim & Kim, 2001; Lee, 2012; Webb, Shavelson, & Maddahian, 1983). 다변량일반화가능도이론을 적용한 연구들로는 국가수준 학업성취도 평가의 분할점수에 대한 오차요인과 학교급과 성취수준별에 따른 최적의 측정 조건 탐색(Kim, Song, & Park, 2012), 미국의 수업 관찰 평가에서 최적의 측정 조건 탐색(Kim, 2014a; Kim, 2014b; Lee et al, 2015; Lee & Han, 2017; Lee & Shin, 2004; Wilhelm & Kim, 2015), 검사에서의 효율적인 측정 조건 탐색(Kim & Choi, 2016; Song & Kim, 2012), 다집단을 고려한 척도의 동등성 분석(Kim, 2017; Kim & Chon, 2018), 합성점수 산출 시 최적의 가중치 탐색(Kim & Han, 2014; Kim & Berebitsky, 2016; Lee, 2012) 등이 있다. 그러나 자동문항생성 기반의 평가에서 다변량일반화가능도이론을 적용한 연구는 현재까지 수행되지 않은 실정이다. 따라서 본 연구에서는 학생들마다 응답해야 하는 문항이 다르게 제공되는 자동문항생성 기반의 평가에서 예시 자료와 다양한 설계를 바탕으로 다변량일반화가능도이론을 활용한 신뢰도 및 효율적인 측정 조건을 탐색하는 방법을 제시하고자 한다. 구체적인 연구문제는 다음과 같다.

첫째, 자동문항생성 기반 평가에 적용할 수 있는 다변량일반화가능도이론의 G-연구는 무엇인가?

둘째, 자동문항생성 기반 평가에 적용한 다변량일반화가능도이론의 D-연구에 따른 신뢰도 및 적정 수준의 신뢰도를 얻는 효율적인 측정 조건은 무엇인가?

Ⅱ. 연구 방법

1. 분석 자료

본 연구에서는 자동문항생성 기반 평가의 신뢰도를 탐색하기 위해 예시 자료로 문항모형을 기반으로 학습으로서의 디지털 평가 플랫폼 구현을 위해 특화된 CAFA (Computer Adaptive Formative Assessment) 시스템(Choi, Kim, & Yoon, 2012-2023; Kim, 2023)을 활용해 생성한 수학 진단평가(이하 평가로 지칭함) 문항에 학생들이 응답한 결과를 활용하였다(Kim *et al.*, 2023). 평가 문항은 진단의 목적으로 활용되므로 2015개정 교육과정의 중학교 3학년 수학의 수와 연산, 문자와 식, 함수, 기하, 확률과 통계 영역에서 생성하였으며 Lee, Lee, & Ham (2022)에서 중등 수학교과 기초학력 신장을 위해 과목별 내용요소, 성취기준을 바탕으로 제작한 대표 문항40개 중 Table 1에서 제시한 바와 같이 수와 연산, 문자와 식, 함수, 기하, 확률과 통계 영역에서 각각 3, 4, 2, 2, 3문항으로 총 14문항을 선정하였다. 구체적으로 중학교 수학교사, 고등학교 수학교사, 수학교육 전문가 4인으로 구성된 전문가 협의회를 통해 평가도구의 내용타당도를 확보하였다. 또한 선정된 문항들은 Figure 1에 예시로 제시한 바와 같이 각각 선택형, 진위형, 그리고 괄호형으로 CAFA 시스템을 통해서 42문항을 생성하였으며, 실제 CAFA 시스템을 사용하고 있는 K시에 위치한 한 여자 고등학교의 1학년 63명을 편의표집으로 선정하여 평가를 실시하였다.

2. 분석 방법

본 연구에서는 예시 자료로 자동문항생성 기반의 평가 결과를 활용하여 일반화가능도이론의 관점을 활용하여 신뢰도를 탐색하였다. 구체적으로 전통적인 신뢰도로 활용되고 있는 Cronbach's α를 활용하기 위해 모든 학생들이 동일한 문항에 응답한 것으로 가정하여 일변량일반화가능도이론의 G-연구 설계 및 원자료의 구조를 반영한 3개의 다변량일반화가능도이론의 G-연구 설계를 적용하였으며, 각 설계는 Figure 2와 같다.

자동문항생성 기반으로 생성된 평가는 모든 학생들이 동일한 문항에 응답하는 것이 아니라 학생들마다

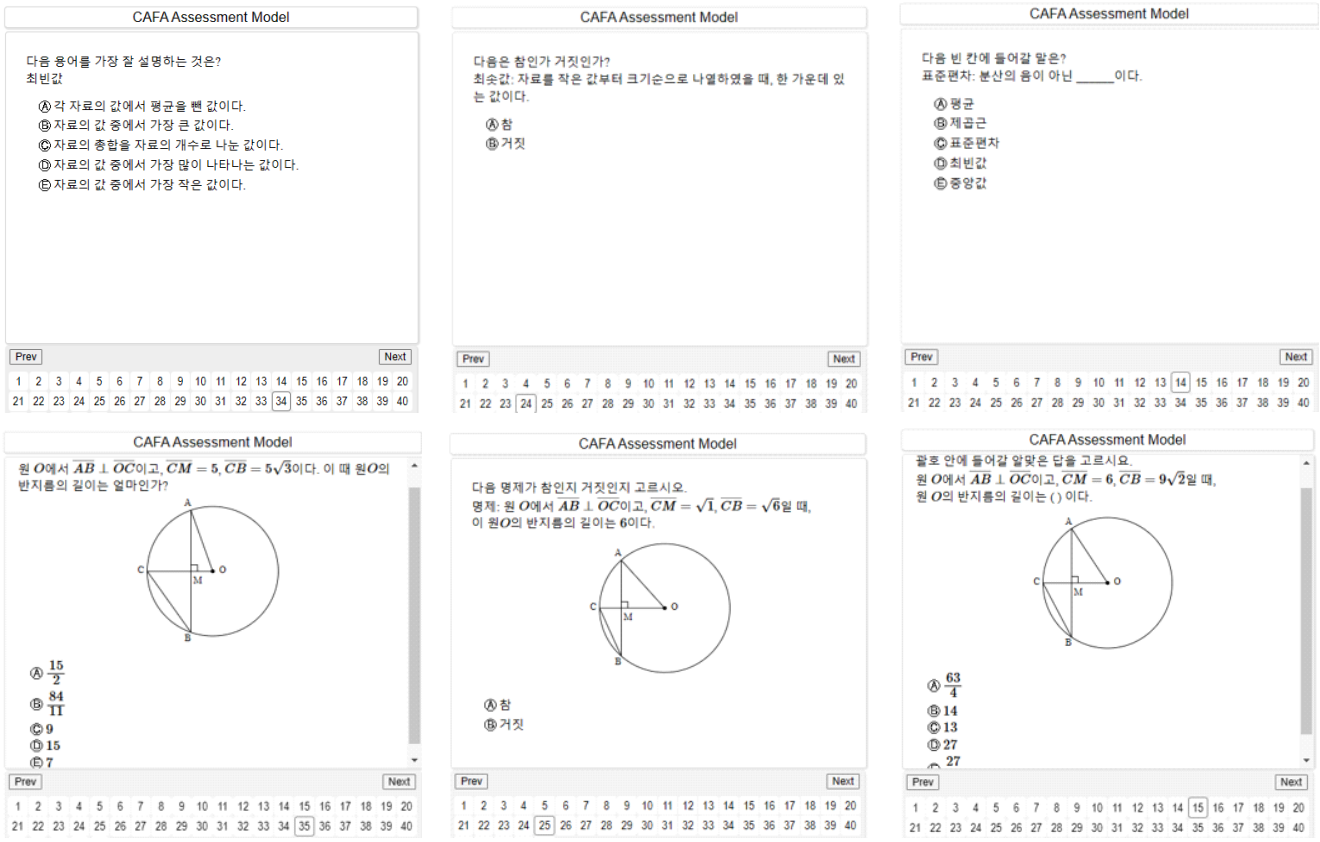

Figure 1. Examples of the diagnostic test for mathematics

다른 문항들에 응답하게 되는 상황이다. 그러나 Figure 1에 제시한 바와 같이 동일한 문항에 대해 유형만 다르게 제공하게 되므로 이러한 자료에 Cronbach's α를 산출하기 위해 일변량 일반화가능도이론의 G-연구 설계로 문항 효과만을 오차요인으로 상정하는 $p \times i$ 설계를 적용하였다. 또한 학생(s)들은 2015개정 교육과정의 중학교 3학년 교육과정에 제시된 수와 연산, 문자와 식, 함수, 기하, 확률과 통계의 고정된 내용 영역(ν)에서 다양한 문항 유형(f) 중 선택형, 진위형, 괄호형에 대해 각기 다른 문항(i)들로 구성된 평가에 응답하게 되므로 이는 다변량일반화가능도이론의 $s^{\bullet} \times (i^{\circ} : f^{\bullet})$ 설계에 해당된다. 다변량일반화가능도이론에서는 고정된 내용 영역별로 먼저 분석이 시행되기 때문에 고정된 국면을 나타내는 ν는 G-연구 설계에 직접 나타내지 않고 Figure 2에 제시된 것처럼 점선으로 표시한다. 또한 G-연구 설계에 닫힌 원(●)은 문항 유형이 고정된 국면인 내용 영역 국면과 교차함을 나타내며, 열린 원(○)은 문항이 고정된 국면인 내용 영역 국면에 내재되어 있음을 의미한다. 유사하게 학생(s)들은 평가에서 고려한 문항 유형(ν)인 선택형, 진위형, 괄호형에서 다양한 수학 내용 영역 중 수와 연산, 문자와 식, 함수, 기하, 확률과 통계에 대해 각기 다른 문항(i)들로 구성된 평가에 응답하는 상황으로 상정할 수 있으므로 이는 다변량일반

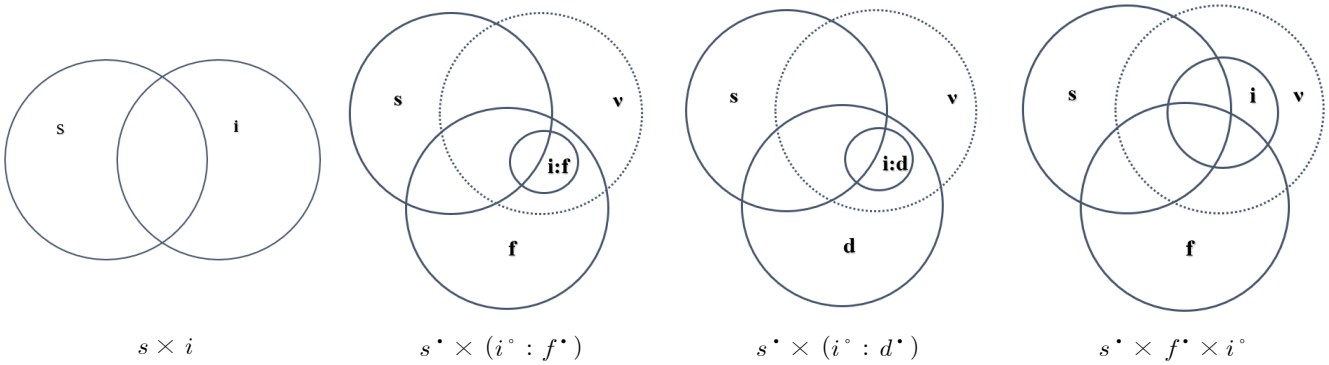

Figure 2. Various G-study designs for a mathematics test based on automatic item generation

화가능도이론의 $s^{\bullet} \times (i^{\circ} : d^{\bullet})$ 설계에 해당된다. 또한 중학교 3학년 교육과정에 제시된 수와 연산, 문자와 식, 함수, 기하, 확률과 통계의 고정된 내용 영역(ν)에서 다양한 문항 유형(f) 중 선택형, 진위형, 괄호형에 대해 모두 같은 문항(i)들로 구성된 평가에 응답하는 것으로 간주하여 다변량일반화가능도이론의 $s^{\bullet} \times f^{\bullet} \times i^{\circ}$ 설계를 적용할 수 있다.

본 연구에서는 편의상 D-연구 설계는 G-연구 설계와 동일하게 적용하였으며, 가장 복잡한 G-연구 설계인 $s^{\bullet} \times f^{\bullet} \times i^{\circ}$ 설계를 바탕으로 효율적인 측정 조건을 제시하였다. 효율적인 측정 조건에 대한 기준으로 적정 수준의 신뢰도(Brennan, 2001a; Dunbar, Koretz, & Hoover, 1991; Shavelson, Baxter, & Gao, 1993; Webb, Shavelson, Maddahian, 1983)보다 엄격하게 적용하여 일반화가능도계수의 경우는 0.85, 그리고 의존도계수의 경우는 0.75로 정하였다. 또한 일변량일반화가능도이론의 G-연구 분석을 위해서는 GENOVA 프로그램을, 다변량일반화가능도이론의 G-연구 분석을 위해서는 mGENOVA 프로그램(Brennan, 2001b)을, 그리고 효율적인 측정 조건의 탐색을 위해서는 엑셀의 매크로를 활용하였다.

Ⅲ. 연구 결과

1. 기술통계 및 G-연구 결과

1) 기술통계

자동문항생성 기반 평가 결과에 대해 문항 유형을 반영한 각 내용 영역별 기술통계 분석 결과는 Table 1과 같다. 확률과 통계 영역의 평균이 가장 높게 나타났으며 함수 영역의 평균이 가장 낮게 나타났다. 문항 유형에 있어서는 확률과 통계 영역을 제외하면 모두 진위형에서 평균이 가장 높게 나타났으며 다음으로 선택형, 그리고 괄호형 순으로 나타났다. 이는 진위형의 경우 응답 범주가 2개로 한정되어 있으며, 같은 문항이어도 어떤 유형으로 제시되느냐에 따라 학생들은 다르게 인식하고 있는 것으로 나타났다.

2) $s^{\bullet} \times (i^{\circ} : f^{\bullet})$ 설계

자동문항생성 기반 평가의 내용 영역을 수와 연산, 문자와 식, 함수, 기하, 확률과 통계로 고정한 $s^{\bullet} \times (i^{\circ} : f^{\bullet})$ 설계의 G-연구 분석 결과는 Table 2와 같다. 구체적으로 Table 2에는 각 효과의 분산과 공분산 성분 추정치, 해당 효과의 분산 성분이 전체 분산에서 차지하는 비율, 그리고 각 내용 영역 간 측정오차를 고려한 수정된 상관계수를 제시하였다.

내용 영역별 분산 성분을 살펴보면 모든 내용 영역에서 잔차 효과가 가장 크게 나타났으며, 확률과 통계 영역을 제외한 나머지 영역에서 고전검사이론의 전집점수에 해당하는 학생 효과가 다음으로 크게 나타났다. 흔히 G-연구 분석 결과에서 잔차 효과에는 G-연구 설계에서 고려한 고차의 상호작용 효과와 본 연구 설계에 포함되지 않은 다른 국면들의 효과가 포함되므로 가장 크게 나타날 수 있다(Brennan, 2001a; Cronbach *et al.*, 1997; Kim, 2014b; Lee *et al.*, 2015; Vallevand, 2008). 또한 고전검사이론의 진점수에 해당하는 학생 효과가 상대적으로 크게 나타남으로써 학생들의 수학 능력 차이가 평가 점수에 반영되어 있다고 해석할 수 있다. 수와 연산, 함수, 기하 영역에서는 문항 유형 차이, 문항 유형에 따른 문항 난이도 차이 순으로, 그리고 문자와 식에서는 문항 유형에 따른 문항 난이도 차이, 문항 유형 차이에 따른 순으로 평가 점수에 영향을 미치는 것으로 나타났다. 또한 모든 내용 영역에서 학생과 문항 유형과의 상호작용 효과가 가장 작게 나타남으로써 학생들의 상대적 순위가 문항 유형에 따라 거의 달라지지는 않는 것으로 나타났다.

Table 2에서 기울임체로 표시한 측정의 오차를 고려한 수정된 상관계수는 0.50845부터 1.0000으로 나타남으로써 한 내용 영역에서 높은 점수를 받은 학생들은 다른 내용 영역에서도 높은 점수를 받고 있으며, 이는 자동문항생성 기반의 평가가 5개 영역으로 구성된 수학 내용을 측정하고 있다는 구인타당도로 해석할 수 있다(Brennan, 2001a; Webb, Shavelson, & Maddahian, 1983). 또한 Brennan (2001a)에 제시된 바와 같이 측정의 오차를 고려한 상관계수가 1보다 큰 경우는 그 값을 1로 표시하였다.

Table 1. Descriptive statistics of a diagnostic test for mathematics

	수와 연산(0.593)			문자와 식(0.608)			함수(0.471)			기하(0.511)			확률과 통계(0.681)		
	선택	진위	괄호	선택	진위	괄호	선택	진위	괄호	선택	진위	괄호	선택	진위	괄호
평균	0.571	0.783	0.423	0.627	0.687	0.512	0.397	0.627	0.389	0.452	0.667	0.412	0.746	0.651	0.646
표준편차	0.346	0.262	0.306	0.265	0.298	0.340	0.393	0.391	0.375	0.356	0.359	0.365	0.266	0.283	0.299
문항 수	3	3	3	4	4	4	2	2	2	2	2	2	3	3	3

주. ( )는 각 영역의 평균을 나타냄.

Table 2. G-study for $s^{\bullet} \times (i^{\circ} : f^{\bullet})$ design

분산 성분	영역	수와 연산	문자와 식	함수	기하	확률과 통계
학생 (s)	수와 연산	0.02521(10.08)	*0.80633*	*0.92595*	*1.00000*	*0.70668*
	문자와 식	0.02906	0.05153(21.01)	*0.98652*	*0.77911*	*0.88476*
	함수	0.03772	0.05745	0.06582(25.06)	*0.76051*	*0.76742*
	기하	0.03483	0.03760	0.04147	0.04519(16.32)	*0.50845*
	확률과 통계	0.01849	0.03310	0.03245	0.01781	0.02717(11.72)
문항 유형 (f)	수와 연산	0.02254(9.01)				
	문자와 식	0.01523	0.00323(1.32)			
	함수	0.02249	0.00952	0.01692(6.44)		
	기하	0.02361	0.01030	0.01818	0.01208(4.36)	
	확률과 통계	0.00000	0.00121	0.00000	0.00000	0.00000(0.00)
문항:문항 유형 ($i:f$)	수와 연산	0.01943(7.77)				
	문자와 식		0.01964(8.01)			
	함수			0.00009(0.03)		
	기하				0.01050(3.79)	
	확률과 통계					0.04495(19.40)
학생×문항 유형 (sf)	수와 연산	0.01185(4.74)				
	문자와 식	0.01122	0.00000(0.00)			
	함수	0.00661	0.00000	0.00000(0.00)		
	기하	0.00990	0.00000	0.01092	0.00000(0.00)	
	확률과 통계	0.00000	0.00000	0.00000	0.00345	0.00000(0.00)
잔차 ($si:f,e$)	수와 연산	0.17105(68.40)				
	문자와 식		0.17083(69.66)			
	함수			0.17981(68.46)		
	기하				0.20908(75.52)	
	확률과 통계					0.15964(68.88)

3) $s^{\bullet} \times (i^{\circ} : d^{\bullet})$ 설계

자동문항생성 기반 평가의 문항 유형을 선택형, 진위형, 괄호형으로 고정한 $s^{\bullet} \times (i^{\circ} : d^{\bullet})$ 설계의 G-연구 분석 결과는 Table 3과 같다. Table 3에는 각 효과의 분산과 공분산 성분 추정치, 해당 효과의 분산 성분이 전체 분산에서 차지하는 비율, 그리고 각 내용 영역 간 측정오차를 고려한 수정된 상관계수를 제시하였다.

문항 유형별 분산 성분을 살펴보면 모든 문항 유형에서 잔차 효과가 가장 크게 나타났으며, 선택형을 제외한 나머지 유형에서 학생 효과가 다음으로 크게 나타났다. 또한 괄호형에서 학생 효과가 가장 크게 나타남으로써 학생들의 수학 능력 차이가 다른 문항 유형보다 괄호형에서 잘 반영되어 있다고 해석할 수 있다. 다음으로 평가에 미치는 영향력으로 선택형과 괄호형에서는 내용 영역에 따른 문항 난이도 차이가, 그리고 진위형에서는 학생에 따라 내용 영역의 난이도 차이를 다르게 인지하는 학생과 내용 영역의 상호작용 효과 순으로 나타났다. 반면에 모든 문항 유형에서 내용 영역 자체의 난이도 효과는 작게 나타났지만, 선택형과 진위형의 경우는 학생들의 순위가 영역에 따라 달라지는 것으로 나타났다. Table 3에서 기울임체로 표시한 측정의 오차를 고려한 수정된 상관계수는 모두 1.0000으로 나타남으로써 한 문항 유형에서 높은 점수를 받은 학생들은 다른 내용 유형에서도 높은 점수를 받는 것으로 나타났다. 따라서 자동문항생성 기반의 평가는 어떤 문항 유형으로 문제가 생성되든지 같은 구인을 측정하고 있다고 해석할 수 있다.

4) $s^{\bullet} \times f^{\bullet} \times i^{\circ}$ 설계

자동문항생성 기반 평가의 내용 영역을 수와 연산, 문자와 식, 함수, 기하, 확률과 통계로 고정한 $s^{\bullet} \times f^{\bullet} \times i^{\circ}$ 설계의 G-연구 분석 결과는 Table 4와 같다. Table 4의 해석은 Table 2와 Table 3과 유사하다. 내용 영역별 분산 성분을 살펴보면 모든 내용 영역에서 잔차 효과가 가장 크게 나타났으며, 고전검

Table 3. G-study for $s^{\bullet} \times (i^{\circ} : d^{\bullet})$ design

분산 성분	문항 유형	선택형	진위형	괄호형
학생 (s)	선택형	0.02521(10.63)	*1.00000*	*1.00000*
	진위형	0.02819	0.01900(8.81)	*1.00000*
	괄호형	0.04275	0.03106	0.05017(19.43)
내용 영역 (d)	선택형	0.00099(0.42)		
	진위형	0.00149	0.00027(0.13)	
	괄호형	0.01341	0.00000	0.00331(1.28)
문항:내용 영역 ($i:d$)	선택형	0.04218(17.78)		
	진위형		0.00568(2.63)	
	괄호형			0.01822(7.06)
학생×내용 영역 (sf)	선택형	0.00882(3.72)		
	진위형	0.00506	0.01629(7.56)	
	괄호형	0.01118	0.01071	0.00000(0.00)
잔차 ($si:d, e$)	선택형	0.16005(67.46)		
	진위형		0.17436(80.87)	
	괄호형			0.18651(72.23)

사이론의 진점수에 해당하는 학생 효과는 수와 연산, 문자와 식, 함수에서는 두번째로, 그리고 수와 연산과 확률과 통계 영역에서는 세번째로 크게 나타났다. 함수 영역에서는 문항 유형의 난이도가, 나머지 영역에서는 문항 유형에 따른 문항의 난이도가 평가 점수에 상대적으로 크게 영향을 미치는 것으로 나타났다. 그러나 문항에 따라 학생들의 상대적인 순위는 거의 변화가 없는 것으로 나타났다. Table 4에 제시된 측정의 오차를 고려한 수정된 상관계수는 0.48388부터 1.0000으로 나타남으로써 Table 2에서와 같이 한 내용 영역에서 높은 점수를 받은 학생들은 다른 내용 영역에서도 높은 점수를 받는 것으로 나타났다.

2. D-연구 결과

Table 5에는 다양한 G-연구 설계들의 비교를 용이하게 하기 위해, G-연구 설계와 동일한 D-연구를 자동문항생성 기반 평가 점수의 평균을 합성점수로 활용하여 분석한 결과를 제시하였다. 구체적으로 고전검사이론의 진점수 분산에 해당하는 전집점수 분산, 규준참조평가에 활용되는 상대오차분산과 신뢰도인 일반화가능도계수, 그리고 준거참조평가에 활용되는 절대오차분산과 신뢰도인 의존도계수를 제시하였다. 또한 모든 학생들이 모두 같은 문항에 응답했다고 가정하여 신뢰도 산출 시 제일 먼저 고려해야 하는 Cronbach α(Churchill, 1979)에 해당하는 일변량일반화가능도이론 설계의 $s \times I$ 분석 결과를 제시하였다. 전집점수분산은 $s \times I$ 설계에서 가장 크게 나타났으며 상대오차분산과 절대오차분산은 가장 작게 나타남으로써 신뢰도인 일반화가능도계수와 의존도계수 모두 가장 높게 나타났다. 이는 자동문항생성 기반 평가 점수에 영향을 미치는 국면을 무시함으로써 신뢰도가 과대추정되었다고 해석할 수 있다.

Table 6과 Table 7은 지면 제약 상 앞서 제시한 G-연구 설계 중 가장 복잡한 형태인 $s^{\bullet} \times f^{\bullet} \times i^{\circ}$ 설계를 바탕으로 적정 수준 이상의 신뢰도에 도달하는 측정 조건들과 합성점수 산출 시 효율적인 가중치를 제시하기 위해 수행한 D-연구 분석 결과를 제시하였다. 구체적으로 Table 6에서 굵은 글자체로 표시한 일반화가능도계수와 의존도계수는 원자료인 평가에서의 신뢰도로 각각 규준참조평가와 준거참조평가에서의 적정 수준 신뢰도인 0.8과 0.7 이상인 것으로 나타났다. 내용 영역은 원자료와 동일하게 5개 영역으로 유지한 상태에서 전체 문항 수를 줄인 결과, 수와 연산과 확률과 통계 영역은 3문항에서 2문항으로 줄이고, 함수 영역은 2문항에서 3문항으로 늘린 경우 전체 문항 수는 39개로 줄었지만, 의존도 계수는 0.80881에서 기울임체로 표시한 0.82065로 높게 나타났다. 또한 문항 수를 원자료와 동일하게 42개로 유지한 경우에도 수와 연산과 확률과 통계 영역의 문항 수는 줄이고, 함수 영역의 문항 수를 늘린 경우에 밑줄로 표시한 것처럼 일반화가능도계수는 0.8729에서 0.90153으로 높게 나타났다. 문항 수를 늘리는 경우 대부분 일반화가능도계수와 의존도계수가 높게 나타났지만 각 영역별 문항 수를 2에서 4까지 차례로 증가시켜본 결과, 총 문항 수가 60개인 경우보다 수와 연산 영역의 문항 수는 줄이고, 함수 영역은 2문항, 그리고 확률과 통계 영역은 1문항 늘린 총 48개 문항의 경우에 가장 높은 신뢰도인 일반화가능도계수는 0.91216, 그리고 의존도계수는 0.85151까지 증가하는 것으로 나타났

Table 4. G-study for $s^{\bullet} \times f^{\bullet} \times i^{\circ}$ design

분산 성분	영역	수와 연산	문자와 식	함수	기하	확률과 통계
학생 (s)	수와 연산	0.02527(9.79)	*0.79098*	*0.86505*	*1.00000*	*0.66948*
	문자와 식	0.02906	0.05342(20.98)	*0.90631*	*0.76776*	*0.82424*
	함수	0.03772	0.05745	0.07523(25.49)	*0.71373*	*0.68086*
	기하	0.03483	0.03760	0.04147	0.04489(15.86)	*0.48388*
	확률과 통계	0.01849	0.03310	0.03245	0.01781	0.03020(12.09)
문항 유형 (f)	수와 연산	0.02254(8.73)				
	문자와 식	0.01523	0.00323(1.27)			
	함수	0.02249	0.00952	0.01839(6.23)		
	기하	0.02361	0.01030	0.01818	0.00913(3.23)	
	확률과 통계	0.00000	0.00121	0.00000	0.00000	0.00000(0.00)
문항 (i)	수와 연산	0.00000(0.00)				
	문자와 식		0.00358(1.41)			
	함수			0.00294(1.00)		
	기하				0.00358(1.26)	
	확률과 통계					0.00358(1.43)
학생×문항 유형 (sf)	수와 연산	0.01185(4.59)				
	문자와 식	0.01122	0.00000(0.00)			
	함수	0.00661	0.00000	0.00000(0.00)		
	기하	0.00990	0.00000	0.01092	0.00000(0.00)	
	확률과 통계	0.00000	0.00000	0.00000	0.00345	0.00000(0.00)
학생×문항 (si)	수와 연산	0.00000				
	문자와 식		0.00000			
	함수			0.00000		
	기하				0.00060	
	확률과 통계					0.00000
문항 유형×문항 (ti)	수와 연산	0.02718(10.53)				
	문자와 식		0.01606(6.31)			
	함수			0.00000(0.00)		
	기하				0.01639(5.79)	
	확률과 통계					0.04732(18.94)
잔차 (sti, e)	수와 연산	0.17123(66.35)				
	문자와 식		0.17839(70.04)			
	함수			0.19863(67.29)		
	기하				0.20848(73.65)	
	확률과 통계					0.16873(67.54)

주1. 기울임체는 측정의 오차를 고려한 수정된 상관계수를 나타냄.
주2. ( )는 영역별 분산 성분이 전체 분산에서 차지하는 백분율을 나타냄.

Table 5. D-study for various designs

분산 성분	$s \times I$	$s^{\bullet} \times (I^{\circ} : F^{\bullet})$	$s^{\bullet} \times (I^{\circ} : D^{\bullet})$	$s^{\bullet} \times F^{\bullet} \times I^{\circ}$
전집점수분산	0.04161	0.03425	0.03516	0.03564
상대오차분산	0.00386	0.00602	0.00448	0.00489
절대오차분산	0.00466	0.00729	0.00788	0.00843
일반화가능도계수	0.91516	0.85062	0.88709	0.87929
의존도계수	0.89937	0.82443	0.81687	0.80881

Table 6. Optimal measurement conditions based on reliability index

영역별 문항 수	전집점수분산	상대오차분산	절대오차분산	일반화가능도계수	의존도계수
(2, 4, 4, 2, 4 : 48)	0.03887	0.00374	0.00677	0.91216	0.85151
(2, 4, 4, 4, 4 : 54)	0.03802	0.00368	0.00685	0.91162	0.84731
(2, 4, 4, 3, 4 : 51)	0.03837	0.00372	0.00683	0.91153	0.84886
⋮	⋮	⋮	⋮	⋮	⋮
(2, 4, 4, 2, 2 : 42)	0.04209	0.00459	0.00849	0.90153	0.83204
⋮	⋮	⋮	⋮	⋮	⋮
(2, 4, 3, 2, 2 : 39)	0.04034	0.00501	0.00881	0.88952	*0.82065*
⋮	⋮	⋮	⋮	⋮	⋮
(3, 4, 2, 2, 3 : 42)	0.03563	0.00489	0.00842	**0.87929**	**0.80881**

다. 이는 문자와 식 영역이 상대적으로 전집점수분산, 상대오차분산, 그리고 절대오차분산이 크게 나타났기 때문이라고 해석할 수 있다.

Figure 3은 문항 효과를 탐색하기 위해 해당 영역을 제외한 나머지 문항 수는 원자료와 동일하게 한 상태에서 해당 영역의 문항 수를 1부터 4까지 증가시켰을 때의 신뢰도와 문항 유형 효과를 탐색하기 위해 나머지는 원자료와 동일하게 하고 문항 유형만 1부터 5까지 증가시켰을 때의 신뢰도를 제시하였다. 구체적으로 전집점수 보다 절대오차분산에 포함되는 문항 유형 효과 및 문항 유형과 문항과의 상호작용 효과가 크게 나타난 수와 연산 영역의 경우에는 문항 수가 증가함에 따라 준거참조평가에 활용되는 의존도계수는 낮아지는 것으로 나타났다. 또한 대부분 문항 유형 수가 증가함에 따라 신뢰도는 높아지는 것으로 나타났지만 최소 3개 문항 유형을 제시하는 경우 적정 수준의 신뢰도에 도달하는 것으로 나타났다.

Table 7은 $s^{\bullet} \times f^{\bullet} \times i^{\circ}$ 설계에서 합성점수를 산출하는 데 있어 각 내용 영역 내의 문항 수를 가중치로 반영하는 상대 가중치(relative weights)와 합성점수에서 각 내용 영역이 실제로 얼마나 반영되어 있는지를 나타내는 실질 가중치(effective weights)를 제시하였다. 구체적으로 연구자들은 원자료의 상대 가중치를 수와 연산, 문자와 식, 함수, 기하, 확률과 통계 각각 0.2143, 0.2857, 0.1429, 0.1429, 0.2143으로 상정하였지만 실제 평가 점수에서는 각각 0.1686, 0.3380, 0.1929, 0.1377, 0.1628로 문자와 식과 함수 영역은 높게, 나머지 영역은 낮게 나타났다. 또한 평가 점수에서 연구자가 상정한 상대 가중치를 얻기 위해서는 내용 영역별로 각각 0.2586, 0.2331, 0.1027, 0.1407, 0.2649로 수와 연산과 확률과 통계 영역을 높여야 하는 것으로 나타났다. 각 영역별 상대 가중치를 0.1씩 증가시키면서 일반화가능도계수와 의존도계수를 살펴본 결과, 각 영역별로 0.1, 0.4, 0.2, 0.1, 0.2의 가중치를 할당하는 경우, 원자료의 신뢰도인 0.87929와 0.80881보다 높은 0.89777과 0.84212까지 높게 나타났다.

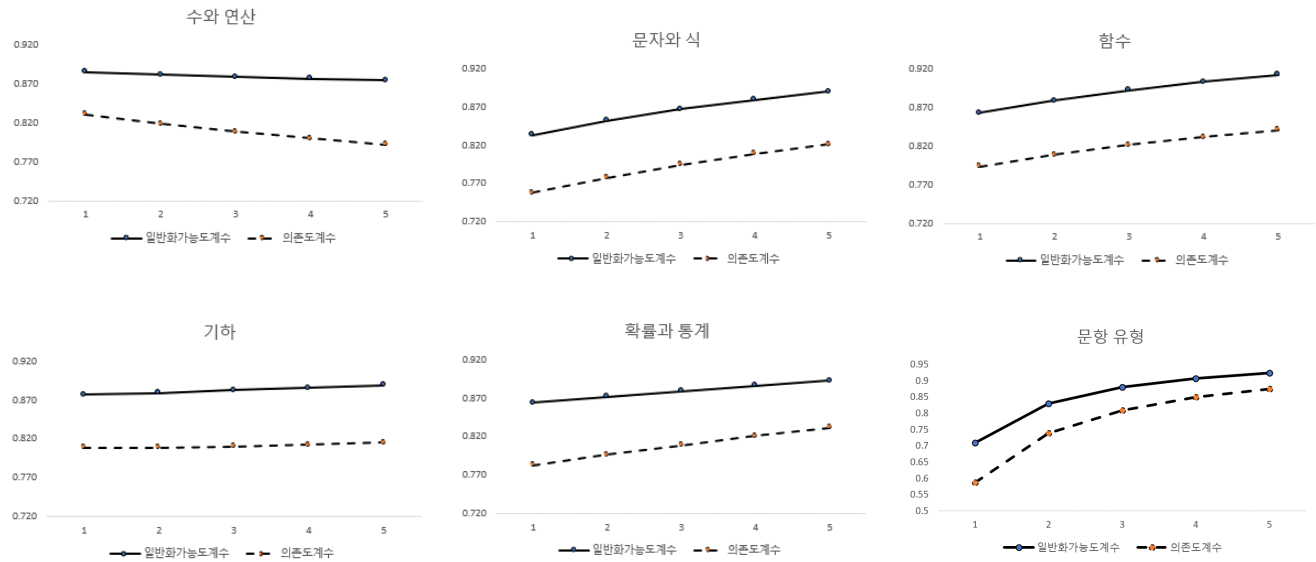

Figure 3. Item effects depending on each domain and Item format effects

Table 7. Optimal relative weights

원자료	상대 가중치					일반화가능도 계수	의존도 계수
	수와 연산	문자와 식	함수	기하	확률과 통계		
	0.1	0.4	0.2	0.1	0.2	0.89777	0.84212
	0.1	0.3	0.2	0.1	0.3	0.89331	0.83981
	0.1	0.3	0.3	0.1	0.2	0.89018	0.83295
수와 연산 : 0.21429	0.1	0.5	0.2	0.1	0.1	0.88919	0.83119
문자와 식 : 0.28571	0.1	0.4	0.3	0.1	0.3	0.88856	0.82820
함수 : 0.14286	0.2	0.3	0.2	0.1	0.2	0.88699	0.81889
기하 : 0.14286	⋮	⋮	⋮	⋮	⋮	⋮	⋮
확률과통계: 0.21429	0. 2586 (0.2143)	0.2331 (0.2857)	0.1027 (0.1429)	0.1407 (0.1429)	0.2649 (0.2142)	0.86266	0.78973
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	(0.1686)	(0.3380)	(0.1929)	(0.1377)	(0.1628)	0.87929	0.80881

주. ( )는 실실 가중치(effective weights)를 나타냄.

Ⅳ. 결론 및 논의

본 연구에서는 일반화가능도이론 관점에서 예시 자료를 활용하여 자동문항생성 기반 평가에서의 신뢰도를 탐색하는 방법을 제안하였다. 주요 분석 결과와 일반화가능도이론이 자동문항생성 기반 평가에 줄 수 있는 시사점에 대해 논의하면 다음과 같다.

첫째, 같은 문항이어도 어떤 유형으로 제시되느냐에 따라 학생들은 다르게 인식하고 있는 것으로 나타났다. 확률과 통계 영역을 제외하면 문항의 난이도는 기존에 가장 많이 사용되고 있는 문항 유형인 선택형보다는 진위형에 정답할 확률이 낮으며 괄호형은 정답할 확률이 높은 것으로 나타났다. 따라서 문항 응답의 선택지가 2개인 진위형의 경우에는 추측에 의해 문항에 정답할 가능성이 높으므로 자동문항생성 기반 평가에서 자동으로 수집할 수 있는 문항을 푸는 데 소요되는 시간 및 학생들이 문항에 응답한 후 돌아가서 재응답을 하는지 등과 관련한 패턴 등을 고려한 연구가 필요하다.

둘째, G-연구 분석 결과 연구 설계와 상관없이 잔차 효과가 평가 점수에 가장 큰 영향을 미치는 것으로 나타났다. 잔차 효과에는 분석에서 고려한 고차의 상호작용 효과가 포함되어있으며, 분석에서 고려하지 못한 다른 국면들의 효과가 포함되어있으므로 일반적으로 크게 나타날 수 있다(Brennan, 2001a; Cronbach *et al.*, 1997; Kim, 2014b; Lee *et al.*, 2015; Vallevand, 2008). 그러나 이는 본 설계에서 고려하지 못한 다른 국면의 필요성에 대한 간접적인 설명으로 해석할 수 있으므로 학생들이 자동문항생성 기반의 평가를 시행한 시간, 장소, 학생들의 특성 및 문항의 특성을 좀 더 세분화한 국면이 고려되어야 함을 알 수 있다. 다음으로 대부분의 설계에서 고전검사이론의 전집점수에 해당하는 학생 효과가 크게 나타났다. 또한 G-연구 설계에 따라 문항 내용 영역을 고정한 설계에서는 문항 유형이나 문항 유형 내 문항에 따라 난이도가 다르게 나타났지만 수와 연산 영역을 제외한 나머지 영역에서 모두 학생들의 상대적 순위는 문항 유형이나 문항에 따라 다르게 나타나지 않는 것으로 나타났다. 문항 유형을 고정한 설계에서도 문항 유형별 점수는 내용 영역에 따른 난이도에는 거의 변화가 없으며, 내용 영역 내 문항에 따른 난이도가 다르게 나타났다. 진위형의 경우 학생들의 상대적 순위가 내용 영역에 따라 달라지는 것으로 나타났지만, 선택형과 괄호형은 변화가 거의 없는 것으로 나타났다. 즉, G-연구 분석 결과 설계에서 고려하는 국면과 설계 자체에 따라 평가 점수에 영향을 미치는 국면의 상대적인 효과가 달라짐을 경험적으로 증명하였다. 이는 자동문항생성을 기반으로 생성된 평가 문항들이 학생들에게 다르게 제시되지만, 문항 유형이나 문항에 따라 학생들의 상대적 순위가 바뀌지 않음을 밝힘으로써 측정학적 특성을 바탕으로 교육 현장에서 자동문항생성 기반의 평가에서 사용하는 서로 다른 문항들의 활용 가능성을 경험적으로 밝혔다.

셋째, D-연구 분석 결과 고전검사이론의 Cronbach α처럼 일변량일반화가능도이론을 바탕으로 하나의 무선 국면을 고려한 설계의 경우에는 오차분산들이 과소추정됨으로써 신뢰도는 과대추정되었다. 또한 고정된 국면의 수준별로 각각 분리하여 분석함으로써 더 정확한 분석 방법으로 알려진 다변량일반화가능도이론(Keller, Clauser, Swanson, 2010)에서도 연구 설계에서 국면을 추가할수록 오차분산이 늘어남으로써 신뢰도는 작아지며, 어떤 국면과 어떤 설계를 활용하는

지에 따라 오차분산과 신뢰도는 달라지는 것으로 나타났다. 그러나 본 연구에서 예시로 활용한 평가의 경우 모든 설계에서 규준참조평가에 활용되는 일반화가능도계수는 0.85이상, 그리고 준거참조평가에 활용되는 의존도계수는 0.75이상으로 나타났다. 또한 적정 수준에 도달하는 효율적인 측정 조건을 탐색한 결과, 문항 수를 원자료보다 줄인 경우에도 신뢰도가 낮은 영역의 문항 수는 줄이고 신뢰도가 높은 영역의 문항 수는 증가함으로써 신뢰도를 각각 일반화가능도계수는 0.879에서 0.890으로, 의존도계수는 0.809에서 0.821로 높일 수 있었으며, 같은 문항 수를 유지한 상태에서도 일반화가능도계수는 0.901까지 의존도계수는 0.832까지 높일 수 있었다. 또한 내용 영역별 가중치를 변화시키면서 효율적인 측정 조건을 탐색한 결과, 원자료와 다르게 가중치를 할당하는 경우, 일반화가능도계수는 0.879에서 0.898로, 의존도계수는 0.809에서 0.842까지 높일 수 있다. 또한 원자료에 상정된 상대가중치가 실제 합성점수에 반영된 정도를 나타내는 실질 가중치는 다르게 나타났으며, 원자료에서 상정된 상대가중치가 실질 가중치에 적용되기 위해서는 각각 내용 영역별로 0.26, 0.23, 0.10, 0.14, 0.27을 할당해야 하는 것으로 나타났다. 이처럼 D-연구 분석 결과는 자동문항생성 기반으로 생성된 평가에서 학생들의 응답 결과에 따라 어느 내용 영역의 문항을 제시해야 할지에 대해 사용될 수 있으며, 안정된 결과를 산출하는 데 있어 더 적은 문항을 제시할 수 있으며, 연구자가 평가도구를 개발하는 데 있어 최초에 상정한 가중치가 실제 평가 점수에서도 반영될 수 있도록 하는 가중치 산출에 활용될 수 있다. 또한 자동문항생성 기반의 평가가 학생들에게 다른 문항을 제공함으로써 측정의 정확성만을 위해 문항을 제시하는 경우, 특정한 분야를 측정하는 문항이 없거나 너무 적어 검사결과의 강건성(robustness)이 줄어들 수 있다는 제한점에 대한 해결 방법으로 평가 개발 내용 영역이나 문항 유형의 균형을 맞추는 데 활용될 수 있다(Choi, 2000).

본 연구의 제한점과 후속 연구를 제안하면 다음과 같다. 첫째, 본 연구에서는 자동문항생성 기반 평가에서의 신뢰도를 탐색하는 방법을 제안하는 것에 초점을 맞추어 예시 평가 자료를 활용하였다. 그러나 예시 자료로 활용된 평가가 개발될 때의 여러 단계에 걸친 타당화 절차에 대한 검증은 이루어지지 않았으므로 이에 대한 검증이 필요하다. 또는 이미 타당도와 신뢰도가 검증된 평가를 활용하여 본 논문에서 제시한 방법론을 적용할 수 있다. 둘째, G-연구 분석 결과에 일반적으로 나타날 수 있는 결과이지만 본 연구에서는 잔차 효과가 가장 크게 나타났다. 이는 본 연구 설계에서 고려하지 못한 다른 요인들의 필요성에 대한 간접적인 설명으로 해석할 수 있다. 따라서 후속 연구로 자동문항생성 기반 평가에서 고려할 수 있는 다양한 요인들을 고려하여 다변량일반화가능도 분석 방법을 활용하는 연구가 수행될 필요가 있다. 셋째, 본 연구에서는 편의상 대부분의 연구에서 D-연구 설계를 G-연구 설계와 동일한 경우에 초점을 맞추어 분석이 수행되었다. D-연구 분석 결과 제시되는 효율적인 측정 조건을 탐색하는 경우 G-연구 설계를 바탕으로 따로 자료를 수집하지 않아도 D-연구 설계를 변화하면서 분석할 수 있다. 따라서 고려할 수 있는 모든 국면을 고려하여 G-연구를 설계하여 교육 현장에서 자료를 수집한 후, D-연구 설계를 다양하게 변화시키는 연구가 수행될 필요가 있다. 마지막으로 본 연구에서는 자동문항생성 기반의 데이터에서 신뢰도를 탐색할 수 있는 방안으로 다변량일반화가능도이론의 적용 부분에 초점을 맞추어 다양한 모형을 제시하였지만, 이 모형 중 어느 모형이 가장 적합한지에 대해서는 논의하지 않았다. 따라서 후속 연구에서는 모형 적합도를 제시하는 구조방정식 모형을 기반으로 다변량일반화가능도 분석을 수행하여 최적의 모형을 선택하는 연구가 수행될 필요가 있다. 또한 본 연구의 결과는 예시 평가에 한정되어 있지만 본 연구에서 활용된 방법론은 자동문항생성 기반으로 수행되는 다양한 평가 분야에 유용하게 적용할 수 있다.

국 문 요 약

본 연구의 목적은 예시 자료를 활용하여 자동문항생성을 기반으로 생성된 평가도구의 신뢰도를 산출하는 방안을 제시하는 데 있다. 이를 위해 학생들마다 다른 문항에 응답하는 설계와 평가 점수에 다중 오차요인을 반영할 수 있는 다변량일반화가능도 이론을 예시 자료에 적용하여 분석하였다. G-연구 분석 결과, 대부분의 설계에서 잔차 효과 다음으로 고전검사이론의 진점수에 해당하는 학생 효과가 크게 나타났다. 또한 문항 내용 영역을 고정한 설계에서 학생들의 상대적 순위는 문항 유형이나 문항에 따라 변하지 않았으며, 문항 유형을 고정한 설계에서 내용 영역에 따라 난이도는 거의 변화가 없는 것으로 나타났다. D-연구 분석 결과, 원자료는 적정 수준 이상의 신뢰도를 확보하였으며, 수와 연산, 기하, 확률 및 통계 영역의 문항 수를 줄이거나 문자와 식과 함수 영역의 가중치를 높게 반영함으로써 원자료보다 높은 신뢰도를 산출할 수 있는 것으로 나타났다. 본 연구에서 제시한 효율적인 측정 조건은 예시 평가 자료에 제한되지만 본 연구에서 활용한 방법은 자동문항생성 기반의 다양한 평가 상황

에서 측정학적 특성을 바탕으로 신뢰도를 산출하고, 효율적인 측정 조건을 탐색하는 데 적용 가능하다.

주제어: 평가, 자동문항생성, 다변량일반화가능도 이론, 신뢰도

References

Brennan, R. L. (2001a). *Generalizability Theory*. New York: Springer.

Brennan, R. L. (2001b). *Manual for mGENOVA* (Version 2.1). [Computer software]. Iowa City, IA: University of Iowa.

Choi, I. (2000). Prospects of computer adaptive testing and performance testing: Language testing in the 21st century. *Language Research, 36*(1), 205-241.

Choi, J. (2020). Educational innovation in the digital age: A plan to implement an intelligent learning analysis platform based on big data. *Educational Development, 214*, 44-50.

Choi, J., Kim, H., & Pak, S. (2018). Evaluation of automatic item generation utilities in formative assessment application for Korean high school students. *Journal of Educational Issues, 4*(1), 68-89.

Choi, J., Kim, S., & Yoon, K. (2012-2023). *CAFA Modeling Manual: Computer Adaptive Formative Assessment User's Guide* [System Manual] (2nd edition). Clarksville, MD: CAFA Lab, Inc.

Choi, J., Oh, K., Youn, K., Lee, D., Joung, J., Kim, S., Youm, S., Lee, Y., Lee, E., Park, W., & Lee, S. (2022). *CAFA model example book.* Seoul: Pubple.

Churchill Jr, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of marketing research, 16*(1), 64-73.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373-399.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-303.

Falcão, F., Pereira, D. M., Gonçalves, N., De Champlain, A., Costa, P., & Pêgo, J. M. (2023). A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation. *Advances in Health Sciences Education,* 1-25.

Gierl, M. J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice, 32*(3), 36-50.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items.* New York: Routledge.

Jeong, H. (2009). A study on computer based test in education environment: Focused on students' experiences. *Journal of Educational Technology, 25*(4), 73-100.

Jeong, J., Shin, K., Lee, S., & Yoo, W. (2009). Design and implementation of iterative contents based on SCORM in mathematics. *Proceeding of the Winter Conference of the Korea Society of Computer and Information, 16*, 153-158.

Kang, S., & Choi, S. (2020). Research on mathematical automatic item generation based on dynamic knowledge-base. *Proceeding of the Korean Information Science Society Conference,* 302-304.

Keller, L. A., Clauser, B. E., & Swanson, D. B. (2010). Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment. *Advances in Health Sciences Education, 15*(5), 717-733.

Kim, S. (2001). An analysis of sources of variation in the observational rating system-comparisons of observer agreement, interrater reliability, and generalizability theory. *Journal of Educational Evaluation, 5*(1), 37-56.

Kim, S. (2014a). Exploring the application of

generalizability theory to mathematics teacher evaluation for professional development in Korea based on the analysis of instructional quality assessment of mathematics teachers in the U.S. *Communications of Mathematical Education, 28*(4), 431-455.

Kim, S. (2014b). Exploring the application of multivariate generalizability theory to teacher evaluation for professional development in Korea based on the analysis of classroom observations in the US. *The Journal of Korean Education, 41*(1), 5-29.

Kim, S. (2017). Multigroup generalizability analysis of Creative Attitude Scale-Korea for mathematically gifted and general students in middle schools. *Communications of Mathematical Education, 31*(1), 49-70.

Kim, S. (2022). The utility of digital evaluation based on automatic item generation in mathematics: focusing on the CAFA system. *The Mathematical Education, 61*(4), 581-596.

Kim, S. (2023). Suggestions on the use of the CAFA system to promote teachers' digital competencies. *The Journal of the Korea Contents Association, 23*(4), 475-493.

Kim, S., & Berebitsky, D. (2016). An application of multivariate generalizability in selection of mathematically gifted students. *Eurasia Journal of Mathematics, Science and Technology Education, 12*(9), 2587-2598.

Kim, S., & Choi, K. H. (2016). An investigation of efficient measurement conditions of the scientific creativity test. *Secondary Education Research, 64*(1), 49-75.

Kim, S., & Chon, K. (2018). An Analysis of measurement errors and invariance properties by proficiency level in the non-cognitive measures. *Journal of Curriculum Evaluation, 21*(1), 153-172.

Kim, S., & Han, K. (2014). Analysis of reliability coefficients depending on different domain weights in scoring teacher recommendation letters and self-introduction letters used in selection of mathematically gifted students. *The Journal of the Korean Society for Gifted and Talented, 13*(1), 43-66.

Kim, S., & Kim, Y. (2001). *Generalizability Theory*. Seoul: Kyoyookbook.

Kim, S., Song, M., & Park, I. (2012). Investigation on optimal conditions and error variance in standard setting using multivariate generalizability analysis. *Journal of Educational Evaluation, 25*(4), 679-700.

Kim, S., Yeum, S., Chung, J., Yoon, K., & Park, S. (2023). Multivariate generalizability theory for reliability with item models: Industrial mathematics test example. *NCME 2023 Annual Meeting*, Chicago, IL: NCME.

Lee, D. (2022, November 22). *4차 산업혁명 시대 따른 미래교육 모습은* [Future education in the era of the 4th industrial revolution]. Retrieved May 01, 2023, from http://m.wsobi.com/news/articleView.html?idxno=182590

Lee, H. (2012). Multivariate generalizability analyses for mixed-format tests with various compositions of MC and CR item weights. *Journal of Educational Evaluation, 25*(1), 95-116.

Lee, J., & Han, K. (2017). Finding optimal condition for conducting TTCT-Figural A (originality) using multivariate generalizability theory. *The Journal of Creativity Education, 17*(2), 57-77.

Lee, J., Lee, S., & Ham, Y. (2022). Case study on college calculus education for vocational high school graduates with coding. Communications of Mathematical Education, 36(4), 611-626.

Lee, S., Kim, S., Kim, J., Baek, K., & Lee, B. (2015). Analyses of the reliability of a preliminary creativity test using the multivariate generalizability theory. *The Journal of Creativity Education, 15*(3), 83-107.

Lee, Y., & Shin, S. (2004). An investigation into the dependability of ratings in a German speaking test using the multivariate generalizability theory. *Foreign Languages Education, 11*(2), 259-265.

Lim, H. (2017). *Study on the design and properties of automatically generated*

items: focusing on polynomial factorization (Unpublished master thesis). Seoul National University, Seoul, Korea.

Ministry of Education [MOE]. (2020). *코로나 이후, 미래교육 10대 정책과제 시안 발표* [Announcement of 10 policy tasks for future education after COVID-19]. Sejong, Korea. Retrieved from https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=294&lev=0&statusYN=W&s=moe&m=020402&opType=N&boardSeq=82145

Ministry of Education [MOE]. (2023). *디지털 기반 교육혁신 방안* [Digital-based education innovation plan]. Sejong, Korea. Retrieved from https://www.moe.go.kr/boardCnts/viewRenew.do?m=060209&s=moe&page=1&boardID=409&boardSeq=94072&lev=0&opType=N

Oh, K. (2022). A study on the development of reading assessment using automatic item generation -Focused on reading comprehension item model-. *Journal of CheongRam Korean Language Education, 87*, 7-34.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

Song, I., & Kim, S. (2012). A validation study of epistemological belief using multivariate generalizability. *The Korean Journal of Educational Methodology Studies, 24*(1), 107-130.

Vallevand, A. L. (2008). *Reliability, validity and sources of errors in assessing physician performance in an Oobjective Structured Clinical Examination: A generalizability theory analysis* (Unpublished doctoral dissertation). Calgary, Alberta, Canada.

Webb, N. M., Shavelson, R. J., & Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Pyans, Jr. (Eds.), *Generalizability theory: Inferences and practical applications* (pp.49-66). San Francisco, CA: Jossey-Bass.

Wilhelm, A. G., & Kim, S. (2015). Generalizing from observations of mathematics teachers' instructional practice using the instructional quality assessment. *Journal for Research in Mathematics Education, 46*(3), 270-279.

저 자 정 보

정 진 민 (아이오와대학교 대학원생)

김 성 연 (인천대학교 교수)