교육과정평가연구
The Journal of Curriculum and Evaluation
2023, Vol. 26, No. 3, pp. 295~318
DOI: https://doi.org/10.29221/jce.2023.26.3.295

LDA 토픽모델링을 활용한 국내 초·중등학생의 자기평가 연구동향 분석[1)]

장윤선 (대구교육대학교 조교수)*

요약

본 연구는 국내에서 수행된 초·중등학생 자기평가 관련 연구의 동향과 주요 주제를 토픽모델링 기법을 적용하여 파악하고자 수행되었다. 이를 위하여 학술연구정보서비스 검색을 통해 논문명 또는 주제어에 자기평가를 포함하며, 초·중등 교실 상황에서 자기평가를 적용하였거나 자기평가 적용을 위한 수업모형 및 프로그램 개발 등 학습자 중심 평가의 일환으로 자기평가를 연구한 논문 113편을 분석자료로 선정하고, 이들 논문의 국문초록을 토픽모델링 기법의 하나인 LDA를 활용하여 분석하였다. 분석 결과 113편의 연구는 크게 세 가지 주제로 분류되었다. 주제별 출현 확률이 높은 단어와 각 주제로 할당된 연구논문 초록의 내용을 바탕으로 '자기평가의 효과성 검증', '자기평가 결과의 일관성 검증', '자기평가에 대한 질적 이해와 활용 방안 모색'으로 각 주제를 명명하였다. 주제별 분포는 '자기평가의 효과성 검증'의 비중이 가장 높았고, 나머지 두 주제의 비중은 비슷하였다. 연구 수행시기에 따른 주제 분포의 변화를 비교하면, 자기평가 관련 연구가 시작된 1990년대 후반에는 '자기평가 결과의 일관성 검증'의 비중이 높았고, 2000년대 이후 '자기평가의 효과성 검증'의 비중이 높아졌으며, 최근에는 '자기평가에 대한 질적 이해와 활용 방안 모색'의 비중이 상대적으로 높았다. 이러한 연구 결과를 토대로 자기평가 연구의 향후 방향과 시사점을 도출하였다.

주제어: 자기평가, 교실평가, 연구동향, 토픽모델링, LDA

1) 이 논문은 2022학년도 대구교육대학교 교내학술연구비 지원으로 수행되었음.
* 제1저자 및 교신저자, ysj@dnue.ac.kr

I. 서 론

최근 교육평가에 대한 패러다임이 '학습 결과에 대한 평가(assessment of learning)'에서 '학습을 위한 평가(assessment for learning)', 더 나아가 '학습으로의 평가(assessment as learning)'로 전환되고 있다. 이러한 교육평가의 패러다임 전환에 따라 우리나라 국가수준 교육과정에서도 과정 중심 평가를 강조하는 것을 비롯하여 학계 및 교육 현장에서 생각하는 교육평가의 기능과 역할이 크게 변화되었다. 구체적으로 학생의 학습 과정에 대한 지속적인 관찰을 통해 학생의 현재 수준과 상황을 파악하는데 필요한 정보를 수집하고 해석하여 학생의 성장과 발달에 필요한 경험이나 정보 등을 제공하는 평가의 형성적 기능이 강조되고 있다(박지현 외, 2020; 박혜영 외, 2018). 새로운 교육평가의 패러다임에서 강조하는 또 하나의 특징은 평가 주체의 다양성이다(박정, 2019; 한국교육과정평가원, 2019). 즉, 평가의 주체를 교사로 한정하는 것에서 벗어나 학생의 주도적이고 적극적인 평가 참여를 강조하는 것이다. 그렇지만 교육 현장을 살펴보면 아직은 평가의 주체를 교사로 한정하는 인식이 강하다. 학습과 평가의 연계를 강조하고 평가의 목적과 기능에 대한 관점이 변화되는 시대적 상황을 고려할 때, 실제 교육과 학습이 이루어지는 교실 안에서 다양한 평가의 효과성을 탐색하고 그 활용도를 제고할 수 있는 방안을 모색할 필요가 있다.

학생 자기평가(self-assessment)는 평가의 주체가 교사가 아닌 학생이 되는 방법으로 단순히 학습의 결과를 평가하는 것에서 그치지 않고, 학습을 계획하고 수행하는 전 과정에 대해 학습자가 스스로 인지하고 이를 평가하는 것을 의미한다. "평가 대상 자신이 주체가 되어 자체의 능력이나 특성을 스스로 진단하고 가치를 부여하는 활동(한국교육평가학회, 2004, p.302)"으로 정의되는 자기평가를 통해 학생은 학습의 주체가 자신임을 인식하고 학습에 적극적으로 참여하게 된다. 더불어 학습에 대한 책임감 증진, 자기주도적 학습 강화 등 자기평가는 여러 측면에서 교육적으로 가치가 높은 평가방법임에도 불구하고 아직 교육현장에서의 활용도는 낮은 편이다. 이는 기본적으로 "자기 자신을 자기가 잘 안다는 전제와 신뢰(한국교육평가학회, 2004, p.303)"를 기본가정으로 하는 자기평가가 실제로는 평가자의 주관을 배제하기 어렵다는 부정적 인식 때문이라 볼 수 있다. 특히, 우리나라에서 실시된 그동안의 교육평가는 학습의 결과를 평가의 핵심 대상으로 여겨왔으며, 학습 결과에 대한 상대적 비교에 치중함에 따라 평가 결과의 객관성을 확보하는 것이 무엇보다 중요하다 생각하는 사회적 분위기로 인해 자기평가를 적극적으로 교실평가에 활용하는 데에는 큰 한계가 있었다. 이러한 이유로 교육평가 분야에서도 자기평가는 상대적으로 많은 관심을 받지 못하였고, 그로 인하여 학생 자기평가를 주제로 수행된 연구논문 편수도 부족한 편이다. 1993년 이후 약 20년 동안 국내에서 발행된 학술논문 가운데 학생 자기평가를 주제로 한 연구논문은 29편에 불과하다(홍소영, 2018).

학생 자기평가는 주관적 판단으로 인해 평가의 신뢰성에 대한 우려가 있으나, 학생이 자신의 인지적 또는 정의적 발달을 위해 필요한 요구에 대한 직접적인 정보를 제공한다는 측면에서 교실평가 맥락에서의 활용도를 제고할 필요가 있다(김동욱, 손원숙, 2020). 이에 본 연구는 새롭게 변화된 교육평가의 패러다임에서 강조하는 학습을 위한 또는 학습으로의 평가와 평가주체의 다양성을 실현할 수 있는 하

나의 방법인 자기평가에 대해 그동안 수행된 연구들을 종합적으로 분석하고, 기존 연구에서 중점을 두었던 연구 주제들이 무엇인지를 탐색하는 과정을 통해 자기평가가 기존 연구자 및 교육 현장에서 어떤 개념으로 인식되고 있으며, 그리고 어떻게 활용되어 왔는지를 파악하고자 한다. 이를 위하여 국내에서 수행된 학생 자기평가에 관련된 연구논문을 수집하고, 수집된 연구논문의 초록을 중심으로 비정형 텍스트 자료를 분석하는 토픽모델링(topic modeling) 기법의 하나인 잠재 디리클레 할당(Latent Dirichlet Allocation: 이하 LDA)을 활용하여 텍스트 자료에 내재된 잠재적인 주제를 도출하는 것을 주요 목적으로 한다. 아울러 분석을 통해 도출된 연구결과를 바탕으로 교육적으로 자기평가의 개념과 정의를 다시 한번 명확하게 정리하고, 교육 현장에서 효과적으로 자기평가 적용의 확대를 도모하기 위하여 향후 필요한 연구 주제는 무엇인지 생각해보는 기회를 제공하고자 한다. 본 연구에서 설정한 주요 연구 문제는 다음과 같다.

첫째, 국내에서 수행된 초·중등학생 자기평가 관련 연구의 현황은 어떠한가?

둘째, 국내에서 수행된 초·중등학생 자기평가 관련 연구의 주요 주제는 무엇인가?

셋째, 국내에서 수행된 초·중등학생 자기평가 관련 연도별 연구 동향의 변화 추이는 어떠한가?

II. 이론적 배경

본 장에서는 이 연구에서 중점적으로 다루고자 하는 자기평가의 개념과 역할 및 기능이 무엇인지, 그리고 실제 교육평가에서 적용되는 자기평가의 방식으로는 어떠한 것이 있는지를 간략하게 살펴보고자 하였다. 더불어 국내 교육평가 관련 분야에서 수행된 연구동향 분석이 어떻게 이루어지고 있는지를 파악하고자 관련 선행연구를 분석하였다.

1. 학생 자기평가

자기평가는 선다형으로 제시된 몇 개의 문항을 통해 자기 스스로 점수를 부여하는 단순한 활동에 국한되는 것이 아니며, 학습이 일어나는 동안 자기 생각과 행동의 질에 대한 모니터링부터 평가에 이르는 전 과정을 포함한다(McMillan & Hearn, 2008). 다시 말해, 자기평가는 학습자가 자신의 학습에 필요한 지식을 이해하고 기술을 습득하는데 필요한 전략이 무엇인지를 파악하는 과정이며, 나아가 향후 이어질 학습의 목표가 무엇인지 알고 목표에 도달하기 위해 필요한 전략에 대한 정보를 제공하는 역할을 한다 할 수 있다. Andrade & Valtcheva(2009)는 자기평가의 형성적 기능을 강조하며, 명확하게 제시된 학습 목표 또는 기준에 근거하여 현재 자기 학습의 질을 파악하고 평가하는 것에서 끝나는 것이 아니라 이를 바탕으로 학습에 사용한 전략이나 수행 또는 학습 결과로 산출된 작품 등을 수정하는 것이 자기평가라 정의하였다. 그러므로 학습과 성취를 복돋우고, 스스로 자신의 학습을 점검하고 관리하는 학업적 자기조절을 촉진하는 것이 자기평가의 주된 목적이라 할 수 있다(Andrade &

Valtcheva, 2009). 이와 같이 정의된 자기평가가 성공적으로 실행된다면 학생들에게 현재 자신이 하고 있는 학습에서 핵심 요소가 무엇인지를 알고 그것에 집중할 수 있게 도와주며, 효과적인 전략과 자신의 강·약점을 파악하는데 도움이 될 수 있다(Andrade & Valtcheva, 2009). 뿐만 아니라 학생들에게 자기주도, 자기반성, 자기결정 및 점검의 경험을 제공함으로써 학습 동기를 부여하고, 능동적인 학습 참여를 유도할 수 있다(손원숙 외, 2015).

교실평가 상황에서 자기평가의 기능이 성공적으로 구현되기 위해 필요한 조건을 Ross(2006)는 다음과 같이 제시하였다. 먼저 학생이 자신의 수행이나 작품을 평가하는데 사용할 기준이 구체적으로 정의되어야 한다. 그리고 이 평가기준을 어떻게 해석하고 사용해야 하는지에 대한 지도가 필요하며, 학생의 자기평가에 대한 피드백이 제공되어야 한다. 마지막으로 자기평가의 결과를 활용하여 학생이 자신의 수행이나 작품을 어떻게 수정하고 향상시킬 수 있는지에 대한 안내도 필요하다. Andrade & Valtcheva(2009)는 Ross(2006)가 제시한 조건들에 덧붙여 실제로 학생이 자기평가 결과를 바탕으로 수행이나 작품을 수정할 수 있는 충분한 시간이 확보되는 것이 중요하며, 자기평가 결과를 최종 성적 산출에는 반영하지 않을 것을 제안하였다.

반면에 자기평가의 방법적 특성으로 인하여 우려되는 측면도 존재하는데. 대표적으로 자기평가 결과에 주관성이 반영될 수 있다는 것이다. 자기평가결과가 최종 성적에 반영된다고 안내를 받은 집단에서 더 관대하게 평가하는 경향이 있으며, 특히 성취수준이 낮은 학생의 경우 실제 성취수준과 자기평가 결과의 차이는 더 크다는 연구(Harris & Brown, 2013; Tejeiro et al., 2012)가 있다. 그 외에도 학생들은 자신의 수행이나 작품을 평가하는 것 활동 자체에 꺼림칙함을 느낄 수 있고(Brown et al., 2009; Ross, 2006), 평가의 목적과 방법에 대한 이해가 부족할 경우 평가활동 자체에 어려움을 겪기도 하며 오히려 그로 인해 실제 학습이나 수행에 집중하지 못할 가능성도 있다. 뿐만 아니라, 자신을 평가하는 과정에서 자신에 대한 부정적인 인식을 형성할 수도 있을 것이다(홍소영, 2018). 하지만 최종 성적에 자기평가 결과를 반영하는 것이 아니라 교수적 피드백 제공을 목적으로 자기평가를 실시했을 때 학생들의 자기평가결과는 교사나 연구자 등 외부 평가자의 결과와 일관된 경향을 보인다는 연구(Bol et al., 2012; Chang. Tseng, & Lou, 2012)도 있다. 마찬가지로 자기평가의 기준을 설정하는 활동에 참여한 학생들은 오히려 자기평가에 대해 긍정적으로 인식하기도 한다(Ross, Rolheiser, & Hogaboam-Gray, 1998).

한편, 선행연구에서 제시된 자기평가의 개념적 정의와 교육적 역할에 대한 논의는 대체로 유사하나 실제 교실평가에서 사용되는 자기평가 활동의 유형과 그 범위는 다양하다(Andrade, 2019; Harris & Brown, 2013). 얼굴 표정이나 신호등과 같은 이미지를 사용하여 학생이 자신의 이해 정도를 표현하는 등의 체크리스트 방법이 가장 단순한 방식이며, 검사를 실시하고 학생이 직접 정답지를 사용하여 자신이 맞힌 문항의 수를 세거나 자신의 수행이나 작품을 수정하기도 한다. 또는 학습한 교과 개념에 대해 이해한(또는 이해하지 못한) 내용을 글로 기술하기도 하며, 좀 더 복잡하게는 채점기준과 예시를 바탕으로 타당한 근거를 들어 자신의 작품이나 수행의 강점과 약점을 파악하거나 학습일지를 작성하는 반성적 활동으로 자기평가를 수행하기도 한다. 이같이 다양한 방식의 자기평가를 Andrade(2019)는 평가의 대상이 무엇인가에 따라 자기효능감 등의 학습 능력(competence), 자기조절 또는 초인지와 같은 학습 과정(process), 발표나 글쓰기와 같은 학습 산출물(product)로 구분하고, 이를 다시 자

기평가의 목적이 형성적 또는 총합적인지, 그리고 학습 과정이나 산출물에 대한 구체적인 평가기준 유무에 따라 총 10가지 유형으로 구분하기도 하였다. 이처럼 다양한 수준과 유형의 활동으로 자기평가를 실시할 수 있으므로 교실평가에서 자기평가가 효과적으로 활용되기 위해서는 각 유형의 평가 방식과 활동이 상황과 맥락에 따라 어떻게 기능하는지에 대한 정보가 수집되어야 하며, 이론적 자기평가의 기능을 경험적으로 검증하는 연구 또한 지속적으로 수행될 필요가 있다.

2. 교육평가 연구동향 분석 연구

다양한 학문 분야에서는 기존에 수행된 선행연구의 현황을 파악하고 그 동향을 분석함으로써 향후 연구의 시사점과 방향성을 제시하는 연구가 수행되고 있다. 교육학의 경우도 교육학과 관련된 여러 분야의 연구동향이 질적 및 양적 분석, 또는 혼합적 분석에 의해 파악되고 있다. 하지만 교육학 내 다른 분야와 비교할 때 교육평가에 대해 수행된 연구의 동향분석은 상대적으로 많이 이루어지지 않았다. 이는 그동안 우리 사회에서 교육평가를 바라보는 관점이 '학습 결과에 대한 평가'로 주로 한정되었고 그로 인해 교육평가 자체를 연구 주제로 삼기보단 학업성취도 분석이나 평가문항 개발 등에 관심이 더 집중되었기 때문이라 생각된다. 그러나 전술한 바와 같이 교육평가에 대한 패러다임이 전환됨에 따라 교실 현장에서의 평가의 역할과 기능이 강조되면서 교육평가 연구의 영역이 교육평가의 방향성, 수행평가나 서논술평가 등의 평가방법, 평가기준 개발, 교사와 예비교사의 평가 역량과 인식 등으로 확대되고 있다. 이러한 맥락에서 최근 국내 교육평가 연구동향을 분석하는 연구가 늘고 있는 추세다.

가장 최근에 수행된 연구로는 2012년부터 2021년 사이의 학술논문 939편을 수집하고, 토픽모델링을 시용하여 국내 학생평가 연구의 동향을 분석한 박민애(2022)의 연구가 있다. 이 연구에서는 최근 10년간 국내에서 수행된 연구들을 10개의 잠재주제로 분류하였는데, 주요 주제로는 교사 및 예비교사의 평가 인식과 전문성, 채점방식과 신뢰도, 교과교육에서의 평가 활용, 과정중심평가와 수행평가, 평가방법에 따른 특성, 평가도구 개발 및 타당화 등이 도출되었다(박민애, 2022). 연구를 통해 최근 국내에서 수행된 학생평가 전반에 관련된 연구 동향을 종합적으로 이해하고, 시기에 따른 주요 연구주제의 변화를 파악하는데 유용한 정보를 제공하였으나, 평가 거버넌스부터 평가 결과의 활용, 교사의 평가 역량에 이르는 다소 광범위한 주제가 포함되어 실제 교육현장에서 활용되는 교육평가가 어떻게 수행되고 그 효과는 어떠한지에 대한 정보를 제공하는데 한계가 있다.

박민애(2022)의 연구가 학생 평가 전반에 대한 연구동향을 파악하는 것에 중점을 두었다면, 일부 연구들은 특정 교과에서의 평가 연구동향을 분석하였다(손태권, 황성환, 2020; 윤문정, 2020; 이건남, 2009; 정수임, 신동희, 2016). 1994년부터 2009년까지 국내에서 수행된 실과교육 평가의 연구를 연구내용과 방법 등을 기준으로 동향을 분석하거나(이건남, 2009), 과학교육 관련 평가 연구를 연구대상, 방법 뿐 아니라 평가 대상 영역을 인지적 영역과 정의적 영역으로 구분하여 연구 동향을 세부적으로 파악하고, 국가 교육과정의 변화 등을 바탕으로 이러한 연구 동향의 요인을 밝히고자 시도하였다(정수임, 신동희, 2016). 윤문정(2020)은 최근 30년 동안 수행된 74편의 연구논문을 대상으로 음악교육에서의 평가 연구 동향을 분석하였으며, 손태권, 황성환(2020)은 국내외에서 수행된 수학교육 평가

연구동향을 토픽모델링을 사용하여 분석하였다.

이 외에 소수지만 특정 평가방법을 대상으로 수행된 연구의 동향을 분석하기도 하였다. 강대중(2018)은 지필평가에서 수행평가로의 변화를 강조한 2009 개정 교육과정에 따라 초등학교 실과교육에서 수행된 수행평가의 계획서 실태를 분석하고, 국내에서 실과교육과 관련하여 수행된 연구논문 465편을 분석하여 실과교과의 수행평가 연구 실태와 동향을 분석하였다. 홍소영(2018)은 1993년에서 2014년까지 수행된 연구논문을 사용하여 학생 자기평가의 학습효과를 일반화하는 연구를 수행하였다. 이를 위해 국내에서 수행된 총 29편의 연구논문에 대한 메타분석을 실시하여 학생 자기평가의 학습효과 크기를 산출하고, 학교급, 교과, 자기평가 방법 등 자기평가의 학습효과와 관련이 있는 요인들의 영향을 종합적으로 탐색하였다. 더불어 분석 결과를 토대로 학생 자기평가가 학생의 학업성취도를 높이는데 긍정적인 효과가 있다는 결론을 도출하였다. 하지만 홍소영(2018)의 연구는 학생 자기평가의 효과성에만 집중함에 따라 자기평가에 대한 기존에 수행된 연구의 동향을 전반적으로 이해하기에는 부족함이 있다.

종합해보면, 교육평가 자체에 초점을 두어 연구동향을 분석한 연구는 최근에 들어서 본격적으로 수행되기 시작하였고, 교과교육에서의 평가에 대한 연구동향을 살펴보는 연구가 주로 이루어지고 있었다. 연구의 수행 연도, 연구대상 및 방법 등을 기준으로 연구의 비중을 비교하는 것이 연구동향을 파악하는 일반적인 방법이며, 최근에 수행된 연구는 토픽모델링을 적용하여 연구동향을 파악하기도 하였다.

III. 연구 방법

본 연구에서는 국내에서 수행된 초·중등학생 자기평가 관련 연구의 동향을 파악하기 토픽모델링을 통해 분석하고자 하였다. 이를 위한 자료 수집과 토픽 모델링 분석을 위해 필요한 텍스트 자료 전처리 및 LDA 분석 절차는 아래와 같다.

1. 자료 수집

한국교육학술정보원에서 제공하는 학술연구정보서비스(Research Information Sharing Service: RISS)를 활용하여 국내에서 수행된 학술논문 및 학위논문을 검색하였다. 연구의 수행 시기는 별도로 설정하지 않았으며, 국내 학술지 논문과 석·박사 학위논문을 대상으로 논문의 제목 또는 주제어에 '자기평가'가 포함된 논문을 검색하여 논문명, 저자, 발행기관, 발행연도, 주제어, 국문초록과 같은 서지정보를 추출하였다. '자기평가'를 검색어로 지정하여 추출된 연구논문은 학술논문 476편, 학위논문 422편으로 총 898편이었으며, 이 중에서 서지정보로 국문초록을 제공하지 않는 학술논문 272편과 학위논문 157편은 분석 자료에서 제외하였다. 이어 국문초록이 존재하는 469편 연구논문의 제목과 국문초록을 검토하여 자기보고식 설문 또는 검사 맥락으로 자기평가가 사용되어 초·중등학생

자기평가와 관련 없는 연구를 분석 대상에서 제외하였으며, 교사 또는 예비교사, 대학생, 성인학습자 등의 자기평가에 관한 연구 역시 초·중등학생의 자기평가 관련 연구의 동향을 파악하고자 하는 본 연구의 목적에서 벗어나므로 분석 대상으로 포함하지 않다. 이상의 기준에 따라 학술논문 36편과 학위논문 78편이 포함된 총 113편이 최종적으로 분석 자료로 선정되었다. 일반적으로 토픽 모델링 연구에서는 논문의 제목, 키워드, 초록과 같은 서지정보를 주요 분석 자료로 사용한다. 본 연구에서는 동일한 단어가 반복적으로 분석에 포함되는 것을 줄이고 연구의 구체적인 내용을 반영하기 위하여 손태권, 황성환(2020)의 연구와 같이 국문초록만 분석 자료로 사용하였다.

연도 및 유형별 연구논문의 형황은 〈표 1〉과 같다. 2000년대 이전에는 '자기평가'를 주제어 또는 논문명에 포함하여 수행된 연구는 4편에 불과하며 모두 석사학위논문이었다. 교육평가에 대한 관점이 '결과에 대한 평가'에서 '학습을 위한 평가'로 전환됨에 따라 2000년 이후 2010년대 중반까지 '자기평가'를 주요 주제로 수행된 연구논문의 수는 크게 증가하였다. 논문의 발행연도를 5년 단위로 나누어 살펴보면, 2010년부터 2014년 동안 수행된 연구논문의 수가 30편으로 가장 많았다. 또한, 석사학위논문 뿐 아니라 박사학위논문과 학술지 게재 논문 등으로 유형이 다양화되었다. 하지만 2014년 이후 5년간의 연구논문의 수는 이전에 비해 다소 감소하였으며, 최근 2022년 이후에는 초·중등학생의 자기평가 관련 연구논문은 검색되지 않았다.

〈표 1〉 연도 및 유형별 연구논문 현황

발행연도	KCI등재	KCI등재후보	박사 학위논문	석사 학위논문	기타*	소계
1995~1999	0	0	0	4	0	4
2000~2004	1	3	0	16	1	21
2005~2009	7	2	0	18	0	27
2010~2014	6	1	2	20	1	30
2015~2019	6	0	2	12	1	21
2020~2021	6	0	0	4	0	10
계	26	6	4	74	3	113

* 학술대회에서 발표된 연구논문은 기타로 분류하였음.

2. 자료 전처리

분석 자료로 선정된 연구논문의 국문초록을 분석하여 잠재된 주제를 도출하기 위하여 텍스트 형태의 자료를 통계적 처리가 가능한 형태로 가공하는 전처리 과정이 필요하다. 전처리의 첫 번째 단계는 분석 자료로 사용되는 국문초록의 문장을 동사, 명사, 형용사, 부사 등의 형태소 단위로 분석하는 것이다. 일반적으로 주제의 의미를 내포하는 명사로 한정하여 주제어로 추출하여 분석에 사용하므로, 본 연구에서도 형태소 분석 후 명사만 추출하였다. 형태소 분석 이후에는 유의어, 불용어 등을 지정하여 분석 자료를 정제하였다. 예컨대, '결핍', '결손', '결여'와 같이 서로 유사한 의미의 단어들은 문자열 치환을 통해 하나로 통일하였고, '서울', '경기도' 등과 같이 특정 지역을 의미하는 단어와 '첫째', '둘째' 등

과 같은 숫자를 표현하는 단어, 한 글자 단어, 그리고 '우리', '각종', 등 주제를 파악하는데 필요한 주요 의미를 가지지 않는 단어는 불용어로 지정하여 분석에서 제외하였다. 또한 분석에 사용되는 모든 연구 논문의 국문초록에 포함되어 문서별 주제를 추출하는데 기여하는 정도가 낮은 '자기평가'도 불용어로 지정하였다. 일부 선행연구에서는 전체 문서에서의 단어 출현 빈도와 개별 문서의 단어 출현 빈도를 고려한 TF-IDF(term frequency-inverse document frequency) 값을 기준으로 1사분위수 이상의 단어만 분석에 포함하거나(손태원, 황성환, 2020), 하위 30개 단어를 불용어로 지정하여 분석에서 제외하기도 하였다(곽민호, 서영진, 2020). 하지만 본 연구에서는 일반적인 텍스트 분석 연구에 비해 다소 적은 수의 문서에서 단어를 추출한 점을 고려하여 전술한 기준에 한정하여 불용어를 지정하였다.

이상의 자료 전처리 과정은 파이썬의 KoNLPy 패키지를 활용하여 수행되었다. KoNLPy은 한글 형태소 분석을 위해 Hannanum, Kkma, Komoran, Mecab, Okt 등 다양한 형태소 분석기를 제공하며, 각 분석기마다 형태소를 분석하는 방식과 수행시간은 조금씩 차이가 있다. 예를 들어 Komoran과 Mecab을 사용하여 형태소를 분석할 경우, 단어를 분절하는 단위가 너무 작아 원래 의미를 반영하는데 부족할 수 있다(김한준, 조새롬, 김동찬, 2021; 박준형, 오효정, 2017). Okt도 마찬가지로 복합단어의 분절단위가 짧은 문제가 있었고, Kkma를 사용하는 경우에는 두 개 이상의 단어로 구성된 복합명사를 구성하는 명사와 함께 기존의 복합명사를 중복적으로 추출하는 문제가 나타났다. 이에 본 연구에서는 각 분석기를 사용한 결과를 비교하여 보다 의미 있는 주제를 추출하기에 적절한 수준으로 문장의 형태소로 분석된 Hannanum의 결과를 최종 형태소 분석기로 사용하였다.

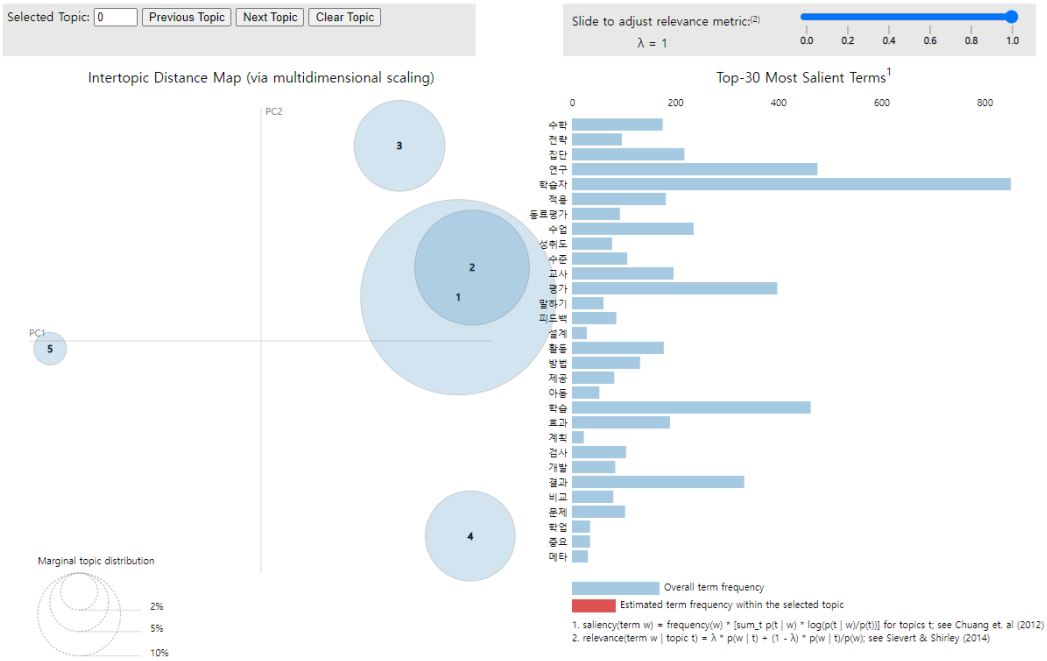
3. LDA 분석

토픽모델링 기법의 하나인 LDA는 다량의 비구조화된 문서집합에 내재된 주제를 추출하기 위하여 디리클래 확률 분포에 기반을 두어 문서집단의 단어 출현 빈도를 분석하는 알고리즘이다(Blei, Ng, & Jordan, 2003). LDA는 분석 자료의 전처리 과정을 통해 생성된 문서집단 내 각 단어의 출현 빈도를 행렬의 형태로 표현한 문서-단어 행렬(document-term matrix)을 사용하여 단어들의 확률적 분포와 관계를 분석하여 서로 관련이 높은 단어의 조합으로 주제를 추출한다. LDA 분석은 파이썬의 Gensim 패키지를 활용하였으며, 이 때 사전에 문서별 주제의 확률과 주제별 단어의 확률 분포를 조절하는 모수를 설정해야 하는데 본 연구에서는 분석에 사용한 Gensim 패키지의 기본 설정을 그대로 사용하였다. 즉, 문서별 주제의 확률 분포를 조절하는 모수(α)와 주제별 단어의 확률 분포를 조절하는 모수(η) 모두'1/주제 수'로 설정하였다(Řehůřek, 2022).

LDA 분석 결과에 대해 타당한 해석을 도출하기 위해서는 최적의 주제 수를 결정하는 것이 중요하다. 주제 수가 너무 작을 경우, 한 주제에 포함되는 단어가 많아지기 때문에 주제의 명확한 의미를 도출하는 것이 어려울 수 있고, 반대로 주제 수가 너무 많은 경우에는 전체 문서와 관련된 주제들 가운데 핵심적이지 않은 내용도 주제로 도출되는 문제가 발생할 수 있다(박준형, 오효정, 2017). LDA 방법에서 주제 수를 결정하는 기준으로 가장 많이 사용되는 것은 혼잡도(perplexity)로 주제가 얼마나 안정적인지를 보여주는 값이다. 주로 혼잡도가 작을수록 적합한 것으로 해석하나(Zhao et al., 2015), 일반

적으로 주제 수가 많아질수록 혼잡도가 감소하는 경향을 보이기 때문에 감소의 정도가 급격해 지는 지점(손태원, 황성관, 2020) 또는 감소의 폭이 완만해지는 지점(이대영, 이현숙, 2021)을 선택하여 주제 수를 결정한다. 또 다른 기준으로 응집도(coherence)를 사용할 수 있다. 응집도는 주제를 구성하는 단어의 연결 정도로 응집도 값이 클수록 각 주제가 잘 구별되고 의미 있게 해석될 수 있음을 의미한다(Du & Liu, 2021). 응집도는 주제 수가 증가 할수록 함께 증가하는 경향이 있으므로 혼잡도를 기준으로 주제 수를 결정할 때와 마찬가지로 변화의 경향을 함께 고려하여 주제 수를 결정하는 것이 일반적이다. 그러나 혼잡도의 결과가 안정적이지 않은 경우가 많으며(Zaho et al., 2015), 혼잡도만을 기준으로 할 때 너무 많은 주제 수가 설정되는 경향이 있다(Gan & Qi, 2021). 응집도 또한 아직까지 주제 수를 결정하는데 있어 신뢰할 만한 결과를 제공하는 것으로 고려하기에는 어려움이 있다(이대영, 이현숙, 2021). 본 연구에서는 주제 수를 2개에서 10개까지 설정하여 각 주제 수에 따라 LDA 분석을 실시한 뒤, Gensim 패키지에서 제공하는 혼잡도와 응집도를 고려하는 동시에 추출된 주제의 비율과 해석의 용이성을 함께 고려하여 주제 수를 결정하였다.

주제의 해석력은 LDAvis 패키지에서 제공하는 시각정보인 주제 간 거리지도(intertopic distance map)와 주제별 단어의 빈도를 참고하였다. 주제 간 거리지도는 [그림 1]의 왼쪽에 제시된 바와 같이 추출된 주제를 이차원 평면에 원형으로 제시함으로써 각 주제에 해당하는 문서의 비율과 주제 간 연관성을 시각적으로 보여준다. [그림 1] 예시의 경우, 하나의 작은 원이 큰 원안에 완전히 겹쳐 제시되었고, 이는 두 주제의 연관성이 상당히 높아 독립적인 주제로 보기 어려운 것으로 해석할 수 있다. 그리고 이차원 평면의 왼편에 가장 크기가 작은 원으로 제시된 주제의 경우 다른 주제들과의 연관성은 낮으나 해당 주제의 비율이 매우 낮음을 보여준다. [그림 1]의 오른쪽에 제시된 정보는 각 주제별 단어의 빈도를 가로 막대그래프 형태로 제공함으로써 각 주제별 주요 단어가 무엇인지를 파악하는데 활용할 수 있다. 이 때, 제공되는 단어는 가중치 모수(λ)에 따라 각 주제 내에서 단어의 출현 빈도가 높은 순으로 제공하거나, 주제 간 출현 빈도의 차이가 큰 단어를 순서대로 제공한다(이대영, 이현숙, 2021). 본 연구에서는 주제 내 단어 출현 빈도를 참고하여 최적의 주제 수를 결정하고, 이후 주제의 의미를 분석하였다. 마지막으로 최종 결정된 주제 수에 따라 추출된 잠재 주제의 의미가 무엇인지 해석해야 한다. 이를 위해 LDAvis에서 제공하는 주제별 출현 빈도가 높은 단어가 무엇인지를 확인하였고, 더불어 각 주제로 할당된 문서의 단어 쌍 동시 출현 빈도를 산출하여 의미연결망 분석을 실시하였다.



[그림 1] LDAvis를 통해 시각화된 LDA 결과 예시

IV. 연구 결과

초·중등학생의 자기평가와 관련된 연구동향 파악을 목적으로 국내에서 수행된 학술지 및 학위논문 113편의 초록을 수집하여 LDA 분석을 실시하였다. 주제 수에 따른 혼잡도와 응집도를 참고로 도출된 잠재주제의 비율과 주제별 주요 출현 단어 등을 종합하여 최적 주제 수를 결정하였고, 각 주제의 주요 단어 및 단어 쌍을 통해 분석한 의미연결망 정보를 종합하여 도출된 주제의 의미를 해석하였다.

1. 주제어 빈도

분석 자료로 선정된 총 113편의 국내 학술지 및 학위논문 초록에서 불용어 제거 등의 전처리 과정을 거쳐 총 1,898개의 단어가 추출되었고, 그 가운데 출현 빈도 기준 상위 30개 단어는 〈표 2〉와 같다. 분석에 사용한 논문의 전체 초록에서 단어의 출현 빈도를 의미하는 TF를 기준으로 가장 많이 등장한 단어는 '학습자'이고, 이어서 '연구', '학습', '평가', '결과' 순으로 출현 빈도가 높았다. 개별 문서에서의 단어 출현 빈도를 고려한 TF-IDF를 기준으로도 '학습자'가 가장 많이 등장하였으나, 그 다음으로는 '학습', '평가', '연구' 순으로 출현 빈도의 순서가 TF를 기준으로 했을 때와 다소 차이가 있었다. TF와

TF-IDF를 기준으로 할 때 모두 출현 빈도 상위 30위에 포함되는 단어는 대체로 동일하였고, 순위만 차이가 있었다.

TF를 기준으로 할 때 상위 30위 안에 포함되었으나 TF-IDF를 기준으로 제외되는 단어는 '방법' '유의미', '변화', '수준'이었고, TF-IDF를 기준으로 할 때에는 이들 단어 대신 '동료평가', '수행', '피드백', '점수'가 포함되었다. 따라서 '방법', '유의미', '변화', '수준'은 자기평가와 관련된 논문 전반에 걸쳐 많이 사용된 단어인 반면, '동료평가'. '수행', '피드백', '점수'와 같은 단어는 일부 논문에서의 사용빈도가 높음을 알 수 있다. 한편, '수학'은 TF를 기준으로 할 때에는 출현 빈도 순위가 15위였으나, TF-IDF를 기준으로 하면 5위로 상승하였다. 마찬가지로 TF 기준 출현 빈도가 23위였던 '영어'도 TF-IDF 기준으로는 11위로 올라오는 것으로 확인되었다. 이는 '수학'이나 '영어'와 같이 특정 교과명은 자기평가를 해당 교과에 적용하여 실시한 논문에서 집중적으로 사용되었기 때문에 TF 기준 순위에 비해 TF-IDF 기준 순위가 상승한 것으로 해석된다. TF 기준과 TF-IDF 기준 출현 빈도 상위 30 단어가 상당부분 중복되었음을 고려할 때 그동안 수행된 자기평가 관련 연구의 주요 내용이나 대상 등에서 큰 차이가 있지는 않은 것으로 파악된다. 다만, 특정 교과에 한정하여 자기평가를 연구한 논문이 일부 존재하며, 자기평가 뿐 아니라 학습자가 평가 주체인 동료평가를 함께 다루거나 피드백을 자기평가의 주요 요소로 다루는 연구도 일부 수행된 것으로 해석된다.

〈표 2〉 113개 초록 내 출현 빈도 상위 30위 단어

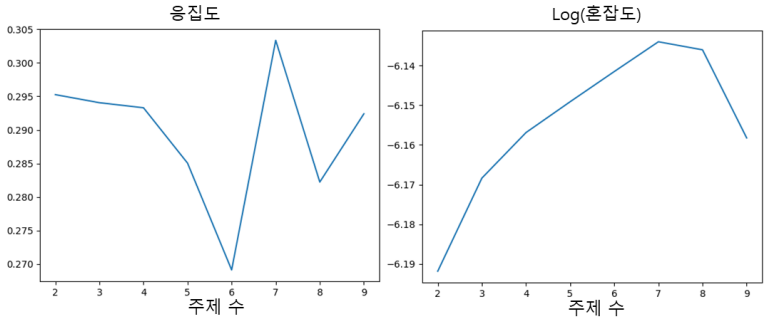
TF 기준						TF-IDF 기준					
순위	단어	빈도	순위	단어	빈도	순위	단어	빈도	순위	단어	빈도
1	학습자	910	16	활동	191	1	학습자	12.77	16	효과	4.17
2	연구	514	17	영향	186	2	학습	8.37	17	적용	3.91
3	학습	491	18	향상	180	3	평가	7.42	18	태도	3.63
4	평가	425	19	활용	179	4	연구	6.73	19	분석	3.62
5	결과	356	20	자신	177	5	수학	6.32	20	동료평가	3.54
6	자기	344	21	학년	163	6	자기	6.16	21	영향	3.52
7	수업	257	22	과정	162	7	수업	5.25	22	자신	3.51
8	실시	253	23	영어	153	8	집단	5.24	23	향상	3.49
9	집단	233	24	태도	153	9	결과	4.93	24	과정	3.26
10	능력	227	25	차이	152	10	능력	4.68	25	활용	3.20
11	교사	215	26	실험대상	149	11	영어	4.68	26	수행	3.16
12	분석	212	27	방법	144	12	활동	4.56	27	피드백	3.14
13	효과	204	28	유의미	125	13	실시	4.52	28	점수	3.13
14	적용	203	29	변화	121	14	교사	4.44	29	차이	3.04
15	수학	192	30	수준	118	15	실험대상	4.21	30	학년	3.03

*TF-IDF 기준에 의해 순위에서 제외 및 추가된 단어는 음영 처리함

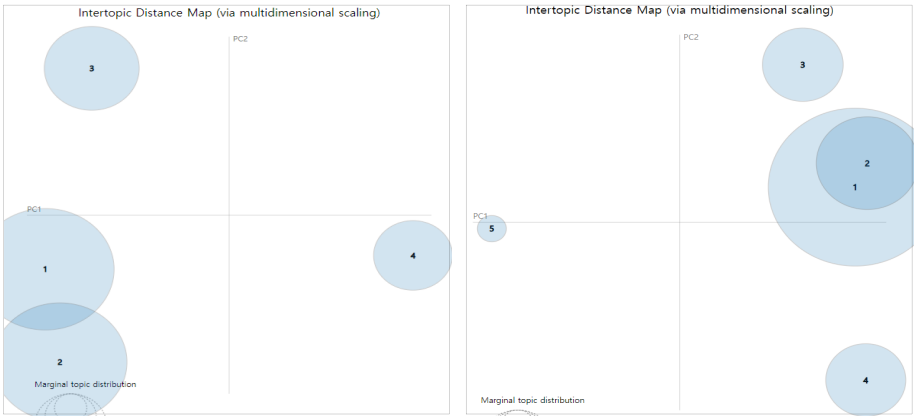
2. 문서 내 잠재주제

최적의 주제 수를 결정하기 위하여 주제의 수를 2개부터 10개로 설정하고 각 주제 수에 따른 모형의 응집도와 혼잡도의 변화 양상을 확인해 보았다. 일반적으로 응집도는 주제 수가 많아질수록 높아지고, 반면에 혼잡도는 낮아지는 양상을 보이나 [그림 2]에 제시된 주제 수에 따른 응집도와 혼잡도의 양상은 이와 다르게 나타났다. 하지만 주제 수에 따른 응집도와 혼잡도의 변화 정도는 실제로 그렇게 크지 않다. 응집도의 경우, 주제 수가 7개일 때 가장 크고 주제 수가 6개 일 때 가장 작으나 이 때 값의 차이는 약 0.03 수준이다. 마찬가지로 혼잡도의 경우도 [그림 2]에 제시된 값이 log(혼잡도)임을 감안하면 실제 혼잡도의 차이는 0.0001 수준에 불과하다. 그러므로 최적의 주제 수를 결정하기 위해서는 단순히 응집도와 혼잡도의 값에 의존하기보다 각 주제 수에 따라 도출되는 주제들의 관계와 비율 등을 종합적으로 고려할 필요가 있다.

[그림 2]의 응집도와 혼잡도 변화 양상을 보면 응집도는 주제 수가 2~4개 사이일 때 대체로 일정 수준을 유지하고 혼잡도는 주제 수가 7개가 되는 시점까지 계속 높아지는 것을 알 수 있다. 또한 본 연구의 분석 자료는 113편 논문의 초록들로 일반적인 토픽모델링 기법을 적용한 연구에 비해 상대적으로 자료의 규모가 크지 않음을 고려할 때, 너무 많은 주제 수를 도출하는 것은 적절하지 않을 수 있으므로 2~5개 사이의 주제 수를 좁혀 도출된 주제를 비교하여 해석적으로 용이한 주제의 수로 결정하는 것이 적절하다 판단하였다. LDAvis를 통해 주제 수를 2~5개로 설정할 때의 LDA 결과를 확인한 결과, [그림 3]과 같이 주제 수가 4개 이상인 경우 주제 간 거리지도에 원형으로 표현된 주제들이 겹쳐 서로 독립적인 주제가 추출되지 않았다. 반면 주제 수가 두개인 경우는 각 주제에 포함되는 단어가 많아짐에 따라 주제에 대한 해석이 다소 모호해 질 수 있다. 따라서 LDA 분석 결과 도출된 주제의 연관성이 낮으며 너무 많거나 작은 수의 문서가 포함되지 않은 3개를 최적의 주제 수로 결정하였다.



[그림 2] 주제 수에 따른 응집도와 혼잡도 양상



[그림 3] 주제 수가 4개와 5개일 때 주제가 거리지도에 표현된 각 잠재주제의 관계

LDA 분석 결과로 도출된 3개의 잠재주제에서 출현 확률이 높은 상위 10개 단어를 제시하면 〈표 3〉과 같다. 문서의 할당 비중이 46.1%(52개)로 가장 높은 주제 1에서는 전체 문서 내에서 가장 많이 등장한 '학습자'의 가장 출현 확률이 높았고, 이어서 '학습', '연구', '수학', '결과' 순으로 출현 확률이 높았다. 그리고 '수학', '실시', '능력', '효과'는 다른 주제에서는 상위 10위 안에 포함되지 않았으나 주제 1에서만 출현 확률이 높은 단으로 포함되어 다른 주제와의 차별되는 의미를 반영하는 것으로 고려할 수 있다. 주제 1로 할당된 연구를 살펴보면, 자기평가가 수학과 학업성취 또는 교과 능력에 미치는 효과를 분석한 실험연구들이 주로 포함되었으며, 그 밖에 과학, 사회, 영어 교과에서의 효과를 분석한 연구도 포함되었다. 또한 일부 연구에서는 인지적 특성 뿐 아니라 학습자의 태도 또는 학업적 흥미 등 정의적 영역에 미치는 효과를 함께 분석하기도 하였다. 따라서 주제 1을 '자기평가의 효과성 검증'으로 명명하였다.

주제 2의 경우, '평가', '학습자', '자기', '연구', '결과' 순으로 주제 내 단어의 출현 확률이 높았고, 다른 주제에서는 출현 확률이 높지 않았으나 주제 2에서 높은 단어는 '자기', '영어', '교사', '점수'였다. 주제 2에는 31개(27.4%) 연구논문이 할당되었는데, 자기평가의 결과와 실제 검사 점수 또는 교사의 평가결과 간 상관을 분석하여 자기평가결과의 타당성을 검증하는 연구가 주로 포함되어 주제 2를 '자기평가 결과의 일관성 검증'으로 명명하였다.

주제 3으로 할당된 문서는 30개(26.5%)로 주제 2와 거의 유사한 비중이었고, 출현 확률이 높은 단어는 '학습자', '연구', '학습', '분석', '수업'이었다. 주제 3에서만 출현 확률이 높은 단어는 '분석', '활동', '반성적', '문제'로 확인되었다. 주제 3으로 할당된 연구를 살펴보면, 수업에 적용한 자기평가 활동에 대한 사례연구와 같은 질적연구와 자기평가를 기반으로 한 수업 설계 또는 평가 프로그램 및 도구 개발 등에 대한 연구들이 많았다. 더불어 자기평가와 공부 또는 심리적 요인 간의 관계를 파악하거나, 실제 학생들이 자기평가를 하는 과정에서 사용하거나 생성된 자료를 질적으로 분석한 연구도 있었다. 따라서 주제 3은 '자기평가에 대한 질적 이해와 활용 방안 모색'으로 명명하였다.

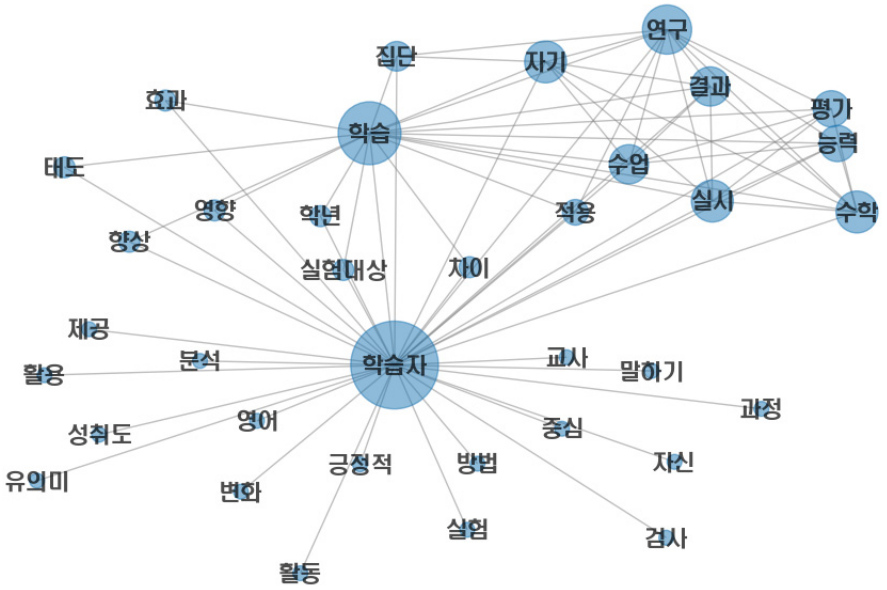
〈표 3〉 주제별 주요 단어 출현 확률(상위 10개)

순위	주제 1	주제 2	주제 3
1	학습자	평가	학습자
2	학습	학습자	연구
3	연구	**자기**	학습
4	**수학**	연구	**분석**
5	결과	결과	수업
6	**실시**	학습	**활동**
7	집단	**영어**	평가
8	수업	**교사**	결과
9	**능력**	**점수**	**반성적**
10	**효과**	집단	**문제**
주제 별 문서 수(비중)	52(46.1%)	31(27.4%)	30(26.5%)

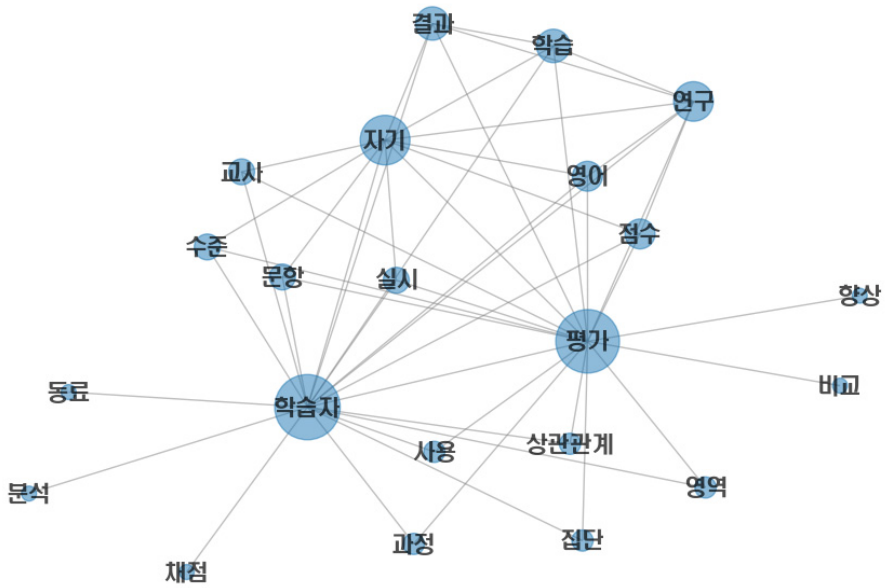
* 주제 간 출현 확률 상위 10개 안에서 중복되지 않은 단어는 진하게 처리함

[그림 4]부터 [그림 6]은 LDA 분석결과 각 주제로 할당된 문서 내 출현 빈도가 높은 단어들이 동시에 출현하는 단어쌍의 빈도를 사용하여 의미연결망을 도식화한 것이다. 동시에 출현하는 단어쌍은 의미연결망에서 노드(node)로 연결되고, 그 출현 빈도는 노드의 크기로 표현된다. 주제별 의미연결망에는 출현 확률 상위 10개보다 많은 단어 간의 관계가 함께 제시됨에 따라 각 주제에 대한 의미를 보다 구체적으로 이해하는데 도움이 될 수 있다. [그림 4]에 제시된 주제 1의 의미연결망에는 '학습자'의 출현 빈도가 높은 동시에 가장 다양한 단어들과 동시에 출현하였음을 보여준다. '학습자'와 동시에 출현하는 단어들로는 주제 1에서 출현 확률이 상위 10위안에 포함되는 '학습', '연구'를 비롯하여 '태도', '효과', '긍정적', '제공', '유의미' 등 이었다. 두 번째로 출현 확률이 높은 단어인 '학습' 또한 출현 확률 상위 10위에 포함된 단어뿐 아니라, '학년', '실험대상', '차이'와 같은 다양한 단어들과 동시에 출현한 것으로 나타났다. 그러나 상위 10위에 포함된 단어 간의 동시 출현 빈도는 대체로 높으나, 그 외 주제 1로 할당된 연구논문의 초록에 등장한 다양한 단어들은 주로 학습자와의 동시 출현 빈도만 높았다.

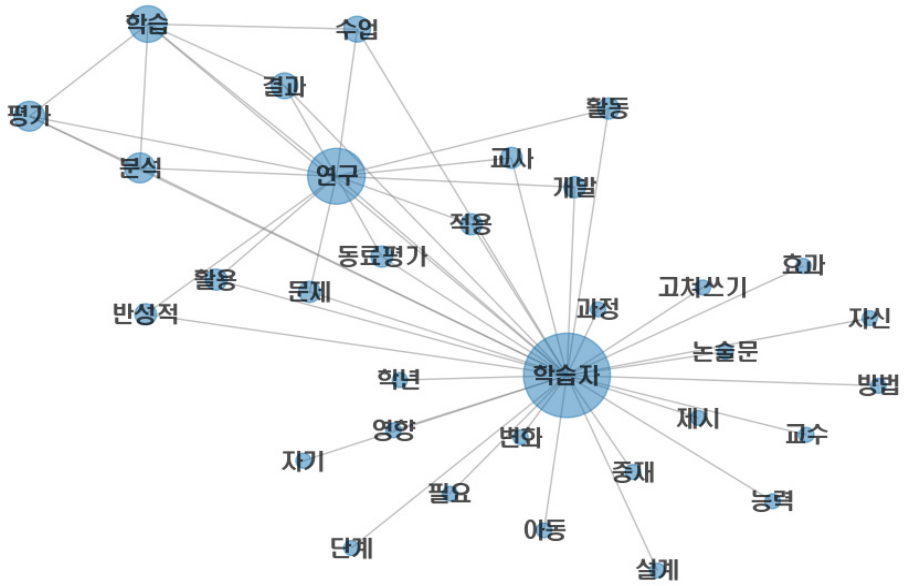
[그림 5]를 살펴보면, 주제 1의 의미연결망과 유사하게 출현 확률 1~3위 단어들은 비교적 다양한 단어들과 동시에 출현하였으나, 그 외 단어들은 주로 출현 빈도가 높은 '평가' 또는 '학습자'와의 출현 빈도만 높았다. 주제 2에서 출현 확률이 가장 높은 '평가'는 주제 2의 주요 출현 단어뿐 아니라 '상관관계', '문항', '과정' 등과 동시에 출현하였고, '학습자'는 주요 출현 단어 이외에 '동료', '채점', '문항' 등과도 연결되어 있었다. [그림 6]에 제시된 주제 3의 의미연결망은 다른 주제에 비해 상대적으로 문서 내 출현 단어들 사이의 연결이 분절적인 경향을 보이며, 대부분 출현 확률이 가장 높은 '학습자'와만 연결되는 단어가 많았다. 이는 주로 주제 3에 할당된 연구 가운데 사례연구의 비중이 높아 각 연구에 한정되는 상황이나 사례 등이 많았기 때문이라 추측된다.



[그림 4] 주제 1'자기평가의 효과성 검증' 의미연결망

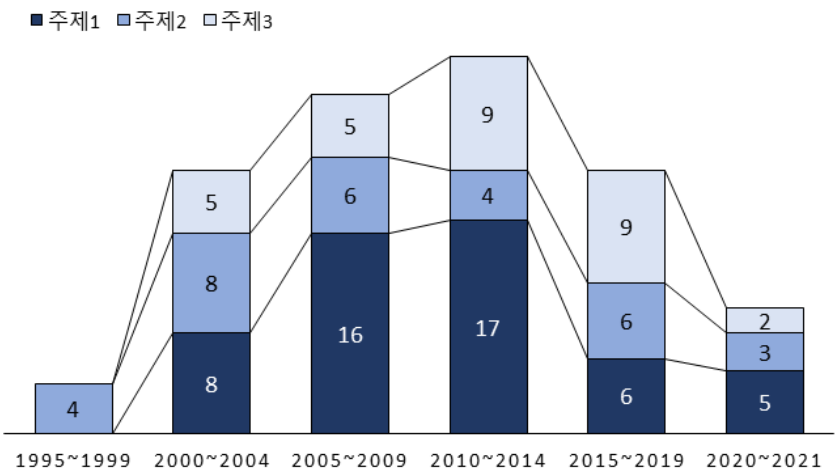


[그림 5] 주제 2'자기평가 결과의 일관성 검증' 의미연결망



[그림 6] 주제 3 '자기평가에 대한 질적 이해와 활용 방안 모색' 의미연결망

마지막으로 본 연구의 분석 대상으로 포함된 113편의 연구가 수행된 연도에 따라 각 주제의 비중이 어떻게 달라지는지를 비교해 보았다. 〈표 1〉과 같이 각 연구의 발행연도를 5년 간격으로 나누어 비교해보면, 자기평가를 주제로 연구가 수행되기 시작한 1995~1999년 사이에는 모든 문서가 '주제 2'로 할당되었다. 즉, 자기평가에 관한 연구가 수행된 초기 단계에서는 자기평가의 결과가 교사 또는 동료 등 다른 평가 주체에 의한 결과와 얼마나 일관되는지를 확인함으로써 평가결과에 평가자의 주관이 개입될 수 있다는 문제에 집중하여 자기평가 적용의 타당성을 검증하는 것이 주요 연구 주제였음을 알 수 있다. 이는 그 당시의 교육평가는 주로 학습 결과를 평가하는 총괄적 성격이 강하였기 때문이라 생각된다. 이후 본격적으로 자기평가를 주제로 하는 연구가 수행된 2000~2014년 사이에는 자기평가가 학생들의 성취도나 능력 등을 향상하는데 얼마나 효과적인지를 양적연구 기반으로 검증하는 주제 1의 비중이 가장 높았다. 학습을 위한 평가를 비롯한 과정 중심 평가 등이 보다 본격적으로 교실평가에 적용되기 시작하는 2015~2019년 사이에는 단순히 자기평가가 대표적인 학습 결과 변인으로 사용되는 학업성취도 또는 능력 등에 어떤 효과가 있는지를 검증하는 주제 1의 연구 비중은 감소하는 반면, 자기평가가 실제 적용된 사례를 중심으로 학생의 학습에 있어 자기평가의 효과와 영향 등을 질적으로 분석하는 주제 3의 비중이 증가하는 것으로 나타났다.



[그림 7] 연도별 각 주제에 할당된 문서 수(개)

V. 결론 및 논의

이 연구는 우리나라에서 수행된 초·중등학생의 자기평가 관련 연구의 동향과 주요 주제를 토픽모델링 기법을 적용하여 파악하는 것을 목적으로 수행되었다. 토픽모델링은 최근 기계학습 기법이 발달함에 따라 많은 양의 텍스트 자료 분석을 통해 연구동향을 파악하는 목적의 연구에서 종종 활용되는 방법으로, 문서집합 내의 단어들의 상관적 확률 분포를 분석하여 상호 간 관계가 강한 단어의 조합에 따라 잠재된 주제를 도출한다. 그러므로 장기간 수행된 많은 양의 연구논문에 내재된 주제를 보다 체계적이고 객관적인 방법으로 분류함으로써 연구동향을 파악하는 토픽모델링 기법을 본 연구에 사용하고자 하였다. 이를 위하여 논문명 또는 주제어에 자기평가를 포함하며, 초·중등 교실 상황에서 자기평가를 적용하였거나 자기평가 적용을 위한 수업모형 및 프로그램 개발 등 학습자 중심 평가의 일환으로 자기평가를 연구한 논문 113편을 분석자료로 선정하고, 선정된 논문의 국문초록을 토픽모델링 기법의 하나인 LDA 방법을 적용하여 분석하였다. 이 연구에서 도출된 주요 결과를 정리하고 그에 대한 논의를 하면 다음과 같다.

첫째, 분석에 사용된 113편의 연구논문은 크게 세 가지 주제로 분류할 수 있었다. LDA 분석을 통해 도출된 세 가지 주제는 '자기평가의 효과성 검증', '자기평가 결과의 일관성 검증', '자기평가에 대한 질적 이해와 활용 방안 모색'이었다. 주제별 문서의 분포는 '자기평가의 효과성 검증'이 52편(46.1%)으로 가장 많았고, 나머지 두 주제는 각각 31편과 30편으로 비슷하였다. 국외에서 수행된 자기평가 관련 연구동향도 이와 대체로 유사한 경향을 보였는데, 2013년부터 2018년 사이 국외에서 수행된 자기평가 연구를 분석한 Andrade(2019)의 연구 결과에 따르면, 76개 선행연구 가운데 자기평가 결과의 일

관성만을 연구 주제로 수행된 경우가 29개로 가장 많았고, 성취도에 대한 자기평가를 핵심 연구 주제로 설정하거나 성취도에 미치는 효과와 함께 자기평가 결과의 일관성 또는 자기조절학습 또는 학습태도, 자기효능감 등과의 관계를 연구 주제로 설정한 연구는 25개인 것으로 확인되었다.

둘째, 도출된 세 주제의 세부 내용을 살펴보면 가장 비중이 높았던 '자기평가의 효과성 검증(주제 1)' 주제로 분류된 연구는 대부분 자기평가 방법이 특정 교과의 성취도나 관련 능력을 높이는데 미치는 영향을 통계적으로 검증하는 준실험연구였다. 그리고 일부 연구에서는 성취도뿐 아니라 학업적 흥미나 태도 등의 학생의 인식에 미치는 효과도 함께 분석하기도 하였다. 이어서 '자기평가 결과의 일관성 검증(주제 2)' 주제에는 대체로 학생의 자기평가 결과와 동료평가 또는 교사평가 간의 상관분석을 통해 그 일관성을 분석하는 연구가 주로 포함되었다. 주제 1 또는 2에는 주로 양적연구가 포함된 것과 달리, '자기평가에 대한 질적 이해와 활용 방안 모색(주제 3)' 주제에는 자기평가 활동을 경험한 학생의 사례연구나 자기평가의 활용가능성 탐색과 같은 질적연구 비중이 상대적으로 높은 것이 특징이다.

셋째, 세 주제의 수행연도를 비교하여 연구 동향의 변화를 파악해본 결과, 자기평가와 관련된 연구가 수행되기 시작한 시기에는 '평가결과의 일관성 검증'과 관련된 연구가 주로 수행되었고, 이후 '자기평가의 효과성 검증' 연구의 비중이 높았고, 최근에는 '자기평가에 대한 질적 이해와 활용 방안 모색' 연구의 비중이 상대적으로 증가하는 경향을 보였다. 이 같은 시기에 따른 연구주제 비중의 변화 경향을 이해하기 위하여 국가 교육과정의 변화를 함께 고려해볼 필요가 있다. 우리나라의 학교 교육은 기본적으로 국가에서 제시하는 교육과정을 기준으로 삼으며, 교실평가 역시 국가 교육과정의 방향에 영향을 받기 마련이다. 자기평가에 대한 연구는 2000년대에 들어서면서 본격적으로 수행된 것으로 확인되었는데, 이는 1997년 12월에 고시된 7차 교육과정 총론에서 처음으로 수행평가 용어가 등장한 것에 영향을 받은 결과로 볼 수 있다. 21세기 지식정보화 사회에 대비하고, 보다 의미있는 평가 결과 제공 등을 목적으로 7차 교육과정에서는 수행평가 도입의 필요성을 강조하였으며, 수행평가의 도입을 통해 선택형 지필검사에 한정한 기존 평가 방식을 다양하게 확대하고자 시도하였다(교육부, 1998). 이러한 국가 교육과정의 방향에 따라 다양한 평가방법에 대한 관심이 높아졌으며, 수행을 평가하는 과정의 일환으로 자기평가가 사용되기 시작함에 따라 자기평가에 대한 연구 수행이 함께 확대되었으리라 생각된다. 하지만 이 시기의 평가는 여전히 결과에 대한 평가를 강조하였으므로 평가결과의 신뢰성 확보가 무엇보다 중요하게 인식되었고, 그로 인하여 자기평가의 일관성을 검증하는 것에 대한 관심 정도가 높았던 것이라 해석된다. 그러나 이후 과정 중심 평가가 교육과정에 명시적으로 등장하며 강조된 2015 교육과정 이후 수행된 연구 가운데 자기평가가 이루어지는 과정 자체에 대한 질적인 이해를 목적으로 하는 연구 수행이 증가하였음을 확인할 수 있었다.

이러한 연구 결과를 종합해 볼 때, 그동안 국내에서 자기평가를 주제로 수행된 연구의 상당수는 주로 단편적인 성취도 향상에 미치는 자기평가의 효과나 다른 방법의 평가결과와 얼마나 유사한 경향이 있는지에 중점을 두었음을 알 수 있다. 그러나 자기평가의 개념적 정의와 교육적 목적의 의미를 깊게 생각해 본다면 학생 평가자의 신뢰도와 자기평가 결과의 정확성, 그리고 자기평가가 학생들의 학업성취도 또는 수행의 향상에 미치는 효과 등을 파악하는 것에 한정하여 자기평가 연구를 수행하고 결과를 도출하는 것은 충분하다 할 수 없다. 특히, 자기평가 결과의 정확성에 대한 연구에서는 대부분 교사와 같은 외부 평가자의 평가는 신뢰할 만한 것이라는 가정을 전제로 자기평가 결과와 외부 평가자의 결과

간의 일치 정도를 분석하지만 실제로 외부 평가자의 평가결과에 대한 타당성을 함께 제시하는 연구는 거의 없다. 그러므로 이러한 연구에서는 자기평가 결과의 정확성(accuracy) 또는 일관성(consistency)이라는 용어를 사용하기보다는 연관성(alignment)이라는 용어가 보다 적절한 표현이라 할 수 있다(Andrade, 2019). 물론 자기평가 결과의 정확성이 향후 학습을 위한 현재 학습이나 수행의 교정에 중요한 역할을 할 것이다. 따라서 향후 연구에서는 단순히 자기평가와 외부 평가자의 결과 간 상관관계를 분석하는 대신 자기평가 결과 활용하여 학생이 자신의 수행이나 작품을 얼마나 정확하게 또는 바람직하게 수정하였는지를 분석한다거나, 자기평가를 정확하게 한 학생과 그렇지 않은 학생의 후속 학습의 결과가 어떠한 차이가 있는지를 비교하는 것으로 연구의 내용 범위를 확장하여 자기평가가 실질적으로 학습과 수행 향상에 어떠한 영향을 주는가에 대한 경험적 정보를 수집할 필요가 있다.

아울러 자기평가가 학습자의 학습과 수행에 어떻게 영향을 미치며, 왜 그러한 영향이 나타나는지에 대한 보다 깊은 이해가 필요하다. 주제 3으로 할당된 김성원(2017)의 연구에서는 중학교 2학년 학생들이 영어교과 활동으로 직접 독해문제를 출제하고 스스로 그 문제를 평가하는 과정을 통해 느낀 긍정적인 측면과 부정적인 측면을 개별 면담을 통해 파악하고, 자기평가 결과와 해당 교과에서의 실제 과업 수행의 관계를 분석함으로써 자기평가가 개별 학습자에게 미치는 영향과 그 원인 및 결과를 심층적으로 이해하고자 시도하였다. 하지만 아직까지 이런 접근의 자기평가 연구는 소수에 불과하므로 향후 지속적인 연구를 통해 경험적 근거를 수집할 필요가 있다. 효과적인 자기평가는 학생이 무엇을 어떻게 평가하는지를 명확히 이해하는 것으로부터 시작된다 할 수 있다. Andrade & Valtcheva(2009)는 단순히 교사에 의해 사전에 준비된 평가기준이 유인물 형태로 제공되기보다 실제 평가기준을 정하는 활동에 학생이 직접 참여함으로써 평가기준에 대한 이해를 높이고 학습을 통해 기대되는 수행이나 작품이 무엇인지를 명확하게 알 수 있도록 하는 것이 보다 효과적일 수 있다고 하였다. 또한, 평가기준이 학습과 관련된 과제의 구체적 특성과 과제를 위해 요구되는 수행을 바탕으로 상세하게 기술된 '수행기반 준거(performance-based criteria)'가 일반적인 '능력기반 준거(competence-based criteria)'에 비해 학생의 학습과 수행을 향상시키는데 더 효과적임을 보여주는 연구(Fastré et al., 2012)도 있다. 따라서 자기평가 과정에서 학생들이 경험하는 여러 활동과 활용하는 자료의 어떤 특성이 학습을 비롯한 인지적·정의적 영역의 긍정적 변화를 가져오는지를 이해하는 것이 향후 자기평가 연구에서 주요하게 다루어져야 할 것이다.

그동안 수행된 국내 연구에서 자기평가는 수업활동 후에 간단한 척도식으로 수행되는 것이 일반적이며, 자기평가 주체의 나이에 따라 학습일지 방식도 많이 활용되기는 하나 학생이 자신의 학습 계획 및 과정에 이르는 전 과정에 대한 반성적 활동이 이루어지지 않은 채 단순한 활동에 대한 느낌이나 감상 수준에 그칠 가능성이 높다는 지적이 있다(홍소영, 2018). 그러므로 자기평가가 실제로 작동하는 원리에 대한 파악과 함께, 이를 반영하여 교육 현장에서 활용할 수 있는 자기평가의 전략이나 활동 또는 평가도구 개발 등에 대한 연구도 함께 지속되어야 할 것이다.

마지막으로 성공적으로 자기평가를 시행하기 위해서는 무엇보다 교사가 자기평가 수행을 위해 학생을 준비시키고 지도함으로써 자기평가로 인해 학생에게 발생할 수 있는 사회적·정서적 문제가 발생하지 않도록 사전에 대처할 수 있어야 한다(Harris & Brown, 2013). 그러나 연구동향 분석 결과에 따르

면 현재까지 수행된 대부분의 자기평가 관련 연구에서는 근본적으로 자기평가에 대한 교사 또는 학생의 이해정도를 조사하거나 분석에 관심을 두지 않았다. 따라서 교사가 인식하는 자기평가의 의미와 교사의 역할이 무엇인지, 그리고 효과적으로 자기평가를 적용할 수 있는 평가 역량을 강화하기 위해 필요한 교육적, 행정적 지원 등이 무엇인지에 대한 연구 등으로 연구주제의 범위가 확대되어야 할 것이다. 더불어 평가에 대한 교사의 책무와 자기평가가 학생 간 또는 학생과 교사의 사회적 관계에 미치는 영향 등 자기평가를 실시할 때 고려해야 하는 다양한 특성에 대한 연구도 추가적인 주제로 제안한다.

그동안 초·중등 학급의 교실평가 맥락에서 수행된 자기평가에 대한 연구의 현황을 종합적으로 이해하고 향후 필요한 연구의 주제와 방향에 대한 기초적인 정보를 제공했다는 측면에서 본 연구의 의의가 있다. 하지만 전술한 바와 같이 일반적인 토픽모델링을 적용한 연구동향 분석에 비해서 본 연구대상에 해당하는 논문의 수가 113편으로 적은 편이고, 그로 인하여 다양한 잠재주제를 도출하지 못하고 한정된 수의 주제가 도출되었다. 도출된 주제 수가 한정적이다 보니 도출된 주제를 기반으로 제시할 수 있는 향후 연구를 위한 시사점에도 제한적이라는 한계가 있다. 이에 국내 학술논문의 서지정보를 제공하는 여러 데이터베이스를 활용하여 가능한 많은 연구 대상을 확보하여 추가적인 분석이 수행될 필요가 있다. 더불어 LDA 방법 외에 다양한 분석방법을 활용하여 자기평가 연구의 동향을 파악하는 것이 필요하다. 예컨대, 초등학생과 중등학생의 발달 수준 차이로 인하여 자기평가의 활동 수준이나 효과가 다르게 나타날 수 있으므로, 연구 대상에 따른 선행연구들을 질적으로 분석함으로써 자기평가에 대한 보다 깊은 이해와 향후 연구의 방향을 제시하는 연구가 수행되기를 기대한다.

참고문헌

강대중(2018). 2009 개정 교육과정 5학년 실과 수행평가 계획서 실태 및 실과 평가 연구 동향 분석. **실과교육연구, 24**(1), 67-82.

곽민호, 서영진 (2020). 잠재디리클레할당(LDA)를 활용한 대학생 동료평가 기준에 관한 탐색적 연구. **교양교육연구, 14**(6), 323-337.

교육부(1998).**수행평가의 이해**. 서울: 교육부.

김동욱, 손원숙(2020). 초등학생용 수업의 질 평가 척도 개발 및 타당화. **교육평가연구, 33**(3), 577-602.

김성원(2017). 학습자 자기평가에 대한 연구: 학습자가 독해문제를 직접 출제하는 학습활동을 통하여. 석사학위 논문, 중앙대학교 교육대학원.

김한준, 조새롬, 김동찬(2021). 경제 텍스트 데이터를 활용한 키워드 분석방안 연구. **한국은행 국민계정리뷰, 1**, 1-41.

박민애(2022). 토픽모델링을 적용한 국내 학생 평가 연구 동향 분석. **교과교육학연구, 26**(2), 155-165.

박정(2019). 과정중심평가를 위한 학생 자기평가 의미 탐색. **교육평가연구, 32**(3), 421440.

박준형, 오효정(2017). 국내 기록관리학 연구동향 분석을 위한 토픽모델링 기법 비교: LDA와 HDP를 중심으로. **한국도서관·정보학회지, 48**(4), 235-258.

박지현, 진경애, 김수진, 이상아(2020). 단위학교의 과정 중심 평가 운영 지표 개발 연구. **교육과정평가연구, 23**(2), 157-181.

박혜영, 이명애, 이명진, 김부연(2018). **미래사회 대비 우리나라 초·중등학교의 교육평가 방향 탐색을 위한 국외 사례 분석**(ORM 2018-39-20). 충북: 한국교육과정평가원.

손원숙, 박정, 강성우, 박찬호, 김경희 역(2015). **교실평가의 원리와 실제: 기준참조수업과의 연계**. 서울: 교육과학사.

Mcmillan, J. H.(2013). *Classroom assessment: Principles and practice for effective standards-based instruction*. MA: Pearson Education, Inc.

손태권, 황성환(2020). 토픽모델링을 활용한 국내외 수학교육 평가 연구 동향 분석. **수학교육학연구, 30**(4), 601-624.

윤문정(2020). 음악교육에서의 평가 연구 동향 분석. **음악교육연구, 49**(4), 153-173.

이건남(2009). 실과 교육평가의 연구동향. **한국실과교육학회지, 22**(2), 1-20.

이대영, 이현숙(2021). LDA 토픽 모델링의 적정 토픽 수 결정 방법 탐색: 혼잡도와 조화평균법 활용을 중심으로. **교육평가연구, 34**(1), 1-30.

정수임, 신동희(2016). 과학 교육에서의 평가 연구 동향. **한국과학교육학회지, 34**(4), 563-579.

한국교육과정평가원(2019). 수업과 연계한 과정 중심 평가, 어떻게 할까요? **현장과 소통하는 KICE 연구·정책 브리프. 14**. 충북: 한국교육과정평가원.

한국교육평가학회(2004). **교육평가 용어사전**. 서울: 학지사.

홍소영(2018). 학생 자기평가의 학습효과에 관한 메타분석. **교육평가연구, 31**(1), 309-331.

Andrade, H. L.(2019). A critical review of research on student self-assessment. *Frontiers in Education. 4*(87), 1-13.

Andrade, H., & Valtcheva, A.(2009). Promoting learning and achievement through self-assessment. *Theory Into Practice, 48*(12), 12-19.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Bol, L., Hacker, D. J., Walck, C. C., & Nunnery, J. A.(2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology, 37*(4), 280-287.

Chang, C. -C., Tseng, K. -H., & Lou, S. -J.(2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. *Computers & Education, 58*(1), 303-320.

Du, B., & Liu, G.(2021). Topic analysis in LDA based on keywords selection. *Journal of Computers, 32*(4), 1-12.

Fastré, G. M. J., van der Klink, M., R., Sluijsmans, D., & van Merriënboer, J. J. G.(2012). Drawing students' attention to relevant assessment criteria: effect on student performance and self-assessment skills and performance. *Journal of Vocational Education & Training, 64*(2), 185-198.

Gan, J., & Qi, Y.(2021). Selecting of the optimal number of topics for LDA topic model: Taking parent policy analysis as an example. *Entropy, 23*(1301), 1-45.

Harris, L. R., & Brown, G. T. L.(2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: Case studies into teachers' implementation. *Teaching and Teacher Education, 23*, 101-111.

Řehůřek. R.(2022). *Gensim: topic modelling for humans*. https://radimrehurek.com/gensim/models/ldamodel.html

Ross, J.(2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research, and Evaluation, 11*(10), 1-12.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A.(1998). Skills training versus action research in-service: Impact on student attitudes to self-evaluation. *Teaching and Teacher Education, 14*(5), 463-477.

Tejeiro, R. A., Gomez-Vallecillo, J. L., Romero, A. F., Pelegrina, M., Wallace, A., & Emberley, E. (2012). Summative self-assessment in higher education: implications of its counting towards the final mark. *Electronic Journal of Research in Educational Psychology, 10*(2), 789-812.

McMillan, J. H., & Hearn, J.(2008). Student self-assessment: The key to stronger student motivation and higher achievement. *Educational Horizons, 87*(1), 40-49.

Zhao, W., Chen, J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W.(2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics, 16*(Suppl 13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8

· 논문접수 : 2023.07.05. / 수정본접수 : 2023.07.28. / 게재승인 : 2023.08.09.

ABSTRACT

Analysing domestic research trends of self-assessment for elementary and secondary school students using LDA based topic modeling

Yoonsun Jang
Daegu National university of Education, Assistant Professor

This study aimed to identify the trends and primary topics in domestic research on self-assessment for elementary and secondary school students using the topic modeling. A total of 113 papers with title or key words related to self-assessment were collected from the Research Information Sharing Service. The Korean abstracts of these papers were analyzed using Latent Dirichlet Allocation(LDA) to discern latent topics. Consequently, the 113 studies were classified into three main topics. namely: 'Analyzing the effectiveness of self-assessment', 'Validation of consistency in self-assessment results', and 'Qualitative understanding and exploring applicability of self-assessment'. The distribution of topics revealed that 'Analyzing the effectiveness of self-assessment' constituted the largest proportion, while the proportions of the other two topics were similar. When comparing the changes in topic distribution over time, in the early stages of self-assessment research, the proportion of 'Validation of consistency in self-assessment results' was higher than those of others. However, as time progressed, the proportion of 'Analyzing the effectiveness of self-assessment' exhibited a gradual increase, and in more recent times, the proportion of 'Qualitative understanding and exploring applicability of self-assessment' was relatively high. Based on these research findings, this study concludes with discussing future directions and implications for self-assessment research.

Key Words: *self-assessment, classroom assessment, research trends, topic modeling, LDA*