English Teaching, Vol. 77, No. 2 Summer 2022, pp. 131-148
DOI: https://doi.org/10.15858/engtea.77.2.202206.131
http://journal.kate.or.kr

영어 읽기 평가에 대한 재고찰: 증거기반평가모델의 적용

소영순*

So, Youngsoon. (2022). English reading comprehension assessment reconsidered: Application of the evidence-centered design. *English Teaching, 77*(2), 131-148.

This paper discusses the limitations of the current practices of English reading assessment in the Korean educational context based on the concept of 'cognitive validity.' It then introduces the Evidence-centered Design (ECD) model as a framework that can guide English teachers and test developers in developing a reading assessment. The paper illustrates how the framework can be applied to the assessment formats and practices widely used in Korean middle and high schools. The ECD framework can help English teachers reconsider reading assessment practices commonly implemented in Korea. The framework contributes to enabling them to focus on the three critical, interrelated questions: what ability to measure with what task(s), how to score students' responses, and how to interpret the test results. Teachers' conscious application of the ECD framework would lead to a more valid and theoretically more sound reading assessment. Such an assessment is expected to align better with teaching and eventually bring a positive washback in English language learning.

Key words: reading assessment, Evidence-centered design (ECD) model, washback/
영어 읽기 평가, 증거기반평가모델, 환류효과

*Author: Youngsoon So, Professor, Department of English Language Education, Seoul National University; 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea; Email: youngsoon_so@snu.ac.kr

Received 6 April 2022; Reviewed 27 April 2022; Accepted 20 June 2022

1. 서론

평가의 핵심은 재고자 하는 지식 혹은 능력, 즉 구인(construct)에 대해 한 학생이 어떤 수준에 있는가에 관한 정보 혹은 증거를 수집하는 것이다. 그리고 평가를 통해 수집된 정보는 특정한 목적으로 사용된다. 예를 들어, 학교에서는 학기 초에 평가를 통해 학생들의 능력을 진단하여 추후 수업의 난이도를 조정하기 위한 목적으로 평가 결과를 활용할 수 있고, 대학에서는 학생 선발의 목적으로 고등학교 내신평가 기록이나 대학수학능력시험 결과를 활용하기도 한다. 따라서, 평가가 목적에 맞는 질 높은 정보를 제공해주는 것은 평가 결과를 기반으로 내린 결정(예. 진단, 선발)의 타당성을 확보하기 위한 기본 전제 조건이 된다고 할 수 있다.

이 논문에서는 '정보 혹은 증거 수집의 수단'으로서의 평가의 본질에 대한 논의를 바탕으로 우리나라 중고등학교 영어 읽기 평가에 대해 비판적으로 고찰해 보고자 한다. 영어의 네 개 기능 중 읽기는 나머지 세 개 기능(듣기, 쓰기, 말하기)에 비해 우리나라 영어 교육 및 평가에서 더 큰 비중으로 다루어진다고 볼 수 있다. 대표적인 예로 대학수학능력시험 (이하 수능) 영어 시험의 전체 45 문항 중 28 개 문항, 즉 약 62%의 문항이 '읽기'와 '간접쓰기'를 포함한 읽기 영역 문항이다(Korea Institute for Curriculum and Evaluation, 2016, p. 1). 따라서, 읽기 기능이 나머지 언어 기능들에 비해 비교적 잘 평가되고 있는 것으로 보일 수 있다. 그러나 이 논문에서는 상대적으로 잘 이루어지고 있는 것으로 보이는 이 읽기 평가를 비판적으로 살펴 보고, 이를 통해 영어 읽기 평가의 타당도를 높이기 위한 방안에 대해 논의하고자 한다. 따라서 이 논문에서 제기하는 핵심 질문은 '우리나라 영어 교육에서 읽기 평가는 제대로 이루어지고 있는가?'이며, 이 질문은 다시 말해 '우리는 영어 읽기 평가를 통해서 학생들의 읽기 능력에 대한 양질의 증거를 수집하고 있는가?'라는 질문으로 바꾸어 쓸 수 있다.

위에서 제기한 문제에 대해 이론적으로 살펴보기 위해서 이 논문에서는 먼저 우리나라 영어 읽기 평가에서 사용되는 '시험 자료', 즉 '읽기 자료' 혹은 '읽기 지문'에 대한 선행연구를 살펴볼 것이다. 이를 바탕으로 그동안 영어 읽기 평가에 관한 논의에서 간과되어 온 이론적인 관점을 평가의 타당도, 그 중에서도 특히 인지타당도(cognitive validity; Field, 2013)라는 개념을 이용해 설명하고자 한다. 이를 바탕으로 평가 문항의 제작부터 평가 결과의 해석에 이르는 일련의 과정에서 평가의 타당도를 고려해볼 수 있는 평가 모델로 '증거기반모델(Evidence-centered design; ECD; Mislevy, Almond, & Lukas, 2003)을 소개하고, 이 모델을 적용하여 우리나라 중고등학교 영어 읽기 평가의 인지타당도 문제에 대해 성찰해 볼 수 있는 기회를 제공하고자 한다. 마지막으로, 본 논문에서 제기된 논의를 바탕으로 평가의 변화를 통해 영어 읽기 교수·학습에 긍정적인 환류효과(washback effect)를 가져올 수 있는 방법을 모색하고 제안해 보고자 한다.

2. 선행 연구

2.1. 읽기 평가에서 제시되는 읽기 지문

읽기 평가(reading comprehension assessment)를 논의할 때 빼놓을 수 없는 구성요소는 읽기 지문(reading passage)과 평가 문항(test item)이다(Alderson, 2000; Hughes, 2003). 이 중 읽기 지문의 특성은 시험을 통해서 수집되는 학생의 읽기 능력에 관한 정보의 질에 큰 영향을 미치게 된다. 이 절에서는 학생들이 이미 접해본 글을 읽기 시험의 지문으로 제시하는 것에 대한 선행 연구를 살펴봄으로써, 학생들이 이미 알고 있는 글을 지문으로 사용한 읽기 평가는 진정한 의미의 읽기 평가라고 볼 수 없다는 점을 지적하고자 한다.

Sung과 Jo(2015)는 서울 시내 8개 학교에서 실시된 고등학교 1, 2학년 영어 내신평가를 분석한 연구로, 우리나라 고등학교 영어 내신평가에 대한 실제적이고 구체적인 정보를 제시하고 있다. 이 연구에 따르면 고등학교 내신평가에서는 읽기 기능과 어휘, 문법 지식의 평가 비중이 쓰기, 읽기, 말하기의 비중에 비하여 훨씬 높았다. 또한, 읽기나 어휘 지식 등을 평가하기 위해 사용하는 지문이 주로 교과서 혹은 수능 모의고사 등에서 이미 학생들이 접해본 지문이라는 결과도 제시되었다. 고등학교 1학년 내신평가 지문의 약 78%, 20%가 각각 교과서와 수능 모의고사 지문을 활용하고 있으며, 고등학교 2학년 내신평가에서는 약 36%, 46%의 지문이 교과서, 수능 모의고사 지문이었다고 연구에서는 보고하고 있다. 다시 말해, 1학년 내신평가 읽기 지문의 98%, 2학년 지문의 약 82%가 학생들이 이미 학습한 지문들로 제시되었다. Sung과 Jo(2015)는 고등학교 내신평가가 문법, 어휘, 읽기 기능 위주로 구성되며, 읽기 평가에서는 학생들이 이미 접했던 지문을 제시한다는 위 연구 결과를 다음과 같이 해석하였다. 고등학교 교사들은 학생들의 수능 준비를 도와줘야 하기 때문에 수능에서 평가되는 읽기 기능이나 어휘, 문법 지식 위주로 내신평가를 구성하게 되고, 문항의 형식이나 언어 자료, 지문도 수능 모의고사를 사용하게 된다는 것이다.

Sung 과 Jo(2015)가 내신평가의 읽기 시험에서 학생들이 이미 학습한 자료를 지문으로 제시하고 있음을 밝혔다면, Lee 와 Lee(2017)는 이미 학습한 자료의 사용이 읽기 시험의 정답률에 미치는 영향을 보여준다. 이 연구에서는 고등학교 교사들이 예측한 읽기 평가 문항 난이도와 학생들의 실제 정답률을 비교하고, 두 결과간 차이가 발견된 경우 그 원인을 알아보기 위해 학생 대상 면담을 실시했다. 이 연구에서 예상 정답률과 실제 정답률 간 차이가 가장 컸던 문항은 '빈칸 추론 문항'이었다. 교사는 이 문항이 가장 어려울 것으로 보고 정답률을 가장 낮게 나올 것으로 예상했으나, 이 문제의 실제 정답률은 예상치보다 훨씬 높았던 것이다. 교사가 해당 문항이 어려울 것이라고 예상했던 이유는 지문 내용의 이해를 넘어 '추론'이라는 상위 인지 과정을 거쳐야 이 문항에 올바르게 답할 수 있을 것이라고 생각했기 때문이다. 그러나 제시된 지문을 이전에 읽어 봐서 내용을 알고 있었던

학생들은 교사가 해당 문항에서 평가하고자 했던 고차원적 인지 과정을 거칠 필요없이 문제를 풀 수 있었다. Lee 와 Lee(2017)가 제시한 다음의 학생 인터뷰는 이러한 해석을 뒷받침해 준다. “저희 부교재에서 거의 똑같이 나온 거라 문제 풀 때 그냥 지문 전체 다 안 읽어도 봤던 지문이니까 답지 보고 바로 정답을 골랐던 것 같아요” (Lee & Lee, 2017, p. 111). 이러한 결과는, 이전에 읽어 봤거나 접해 본 글을 지문으로 제시할 경우 학생들은 출제자가 해당 문항에서 의도한 인지적 과정을 거칠 필요가 없으며, 그 결과 평가하고자 의도했던 인지 과정(예. 추론 능력)에 대해 학생이 갖고 있는 능력 수준에 대한 정보를 얻기 어렵게 된다는 것을 보여주는 좋은 사례이다.

위에서 논의한 Sung 과 Jo(2015), Lee 와 Lee(2017)의 연구는 고등학교 내신평가를 연구 대상으로 삼았으나, 이 논문들을 통해서 우리나라 고등학교의 내신평가가 수능의 영향을 많이 받고 있다는 사실도 함께 확인할 수 있다. 교사들은 내신평가에 수능 문항과 유사한 문항을 출제하며, 읽기 지문으로는 EBS 수능 교재의 지문을 활용하고 있다. 내신평가에 교과서 지문을 사용하는 것의 문제점을 지적한 연구는 이미 1990 년대 중반에도 출판된 적이 있으나(Kahng, 1995), 이런 문제제기에도 불구하고 내신평가뿐 아니라 국가에서 관할하는 시험인 수능에서까지 학생들이 이미 읽어본 지문을 사용해 온 것은 다소 안타까운 일이라고 할 수 있다.

그러나 Kwon(2015)에 따르면 첫 수능이 시행된 1993 년부터 1999 년 초까지만 해도 영어 영역의 지문은 교과서 밖에서 출제하는 것이 원칙이었다. 그러던 중 2000 년대에 들어와 교과서 내용의 출제 비율을 높여야 한다는 논의가 시작되었고, 이후 이러한 논의는 EBS 강의 내용을 수능 시험에 반영하는 것, 더 나아가서는 EBS 교재에 실린 지문을 수정없이 그대로 수능에 출제하는, 소위 직접 연계 정책으로 이어졌다. 수능-EBS 연계 정책은 결국 고등학교 교육과정과 수능 시험 간의 괴리를 키우는 결과를 초래했으며, 수능이 본래 의도대로 고차원적 사고능력을 재는 시험이 아니라 암기력을 평가하는 시험이 되게 했다는 점을 Kwon(2015)은 비판한다. 다행히도 2022 학년도 수능부터는 EBS 교재의 지문을 그대로 사용하는 직접 연계가 아니라 주제, 소재 등에서 유사한 지문을 사용하는 간접 연계로 정책의 변화가 있었으나, 여전히 학생들이 읽어본 내용과 어떤 방식으로든 연관되어 있는 지문이 읽기 시험의 지문으로 제시되고 있다는 점은 변하지 않았다.

위에서 논의한 바와 같이, 우리나라 내신평가와 수능의 읽기 평가에서 학생들이 이미 읽어본 지문을 제시하는 것은 읽기 평가의 이론에 반한다는 점이 본 논문에서 제기하고자 하는 문제의 핵심이다. Hughes(2003)는 읽기 평가의 지문 선택 시 주의사항 10 가지를 제시하고 있다. 그는 이 주의사항들이 “자명해보이지만 자주 간과되는(While the points may seem rather obvious, they are often overlooked.)” (p. 142) 사항들이라고 지적한다. 그가 제시한 10 개 사항 중 다음 사항은 본 논문의 주장과 밀접하게 관련되어 있다.

Hughes 는 “학생들이 이미 읽어본 지문(혹은 그와 아주 유사한 지문)을 읽기 지문으로 제시하지 말라(Do not use texts that students have already read (or even close

approximation to that)"(Hughes, 2003, p. 143)고 제시하며, "놀랍게도 이런 지문이 사용되는 경우가 빈번하다(This happens surprisingly often)"(Hughes, 2003, p. 143)고 덧붙인다. 학생들이 이미 읽어본 지문을 사용하는 것이 비단 우리나라 영어 읽기 평가에서만 나타나는 현상은 아니라는 것을 알 수 있지만, 그것이 그동안 우리나라 영어 읽기 평가에서 큰 문제의식 없이 학생들이 이미 읽었던 글을 시험 지문으로 사용해 왔던 점을 정당화해 주지는 못한다.

그렇다면 왜 학생들이 이미 읽은 지문이나 이와 비슷한 지문조차도 읽기 시험의 지문으로 사용하면 안 되는가? 그 이유는 Lee 와 Lee(2017)의 연구에 참여했던 학생이 인터뷰에서 말했던 내용처럼, 지문을 읽지 않고도 정답을 맞출 수 있기 때문이다. 즉, 학생이 '읽기' 과정을 거치지 않고 문제를 푸는 경우 우리는 그 학생의 '읽기' 능력에 대한 양질의 정보를 수집할 수 없으며, 이는 평가 이론에서 중요한 요소인 타당도, 그 중에서도 인지타당도(cognitive validity)에 부정적인 영향을 미치게 된다. 다음 절에서는 이 인지타당도의 개념에 대해서 조금 더 깊이 논의해보고자 한다.

2.2. 인지타당도(cognitive validity)

평가의 인지타당도(cognitive validity)란 "수험자가 비평가 상황에서 거칠 법한 인지적 과정이 평가에서 요구되는 정도(the extent to which a test requires a candidate to engage in cognitive processes that resemble or parallel those that would be employed in non-test circumstances)"(Field, 2013, p. 78)로 정의된다. 이 개념은 평가의 타당도 논증(validation argument)에 대한 Messick(1989)의 이론과 일맥상통하는 것이다. 평가의 타당도 논증을 위해서는 다각도의 증거를 종합적으로 검토할 필요가 있음을 강조한 이 논문에서 Messick 은 "수험자들이 평가 문항이나 과업을 어떤 방식으로 처리하는지 살펴봄으로써 문항에 대한 응답이나 과업 수행의 기반이 되는 과정을 밝힐 수 있다(We can directly probe the ways in which individuals cope with the items or tasks, in an effort to illuminate the processes underlying item response and task performance)"(Messick, 1989, p. 6)고 강조한다. 즉, 학생들이 시험 문제를 풀기 위해 거치는 인지 과정(cognitive process)이 비시험 언어 사용 상황(non-assessment language use situation)에서 요구되는 과정과 비슷할 경우에, 그 시험을 통해서 수집된 정보를 바탕으로 학생들이 실제 언어 사용 상황에서 그 능력을 어느 정도로 발휘할 수 있는지 정확하게 추론할 수 있다는 것이다.

만약 시험 상황에서 학생들이 시험 문제를 풀기 위해서 거치는 과정이 실제 언어 사용 상황과 다르다면 어떻게 될까? 이러한 경우에는 평가에서 재고자 하는 능력, 즉 평가 구인(construct)이 아닌 다른 요인, 즉 '구인과 무관한 원인(construct-irrelevant source)'이 학생들의 평가 결과에 영향을 미치게 되며, 이는 평가 결과의 타당도(validity)를 낮추는 주요 원인이 된다. 이러한 이유로 새로운 평가의 개발 과정(Cho & So, 2014)이나 이미 개발된 평가의 타당도 검증 연구(Winke et al., 2018)에서는 수험자들이 평가 과업 수행 시 거치는 인지적 과정을 살펴보게 된다.

예를 들어, Cho 와 So(2014)나 Winke 외(2018)는 수험자가 시험 문항에 답을 하는 과정을 면담을 통해 알아보고 있다. 만약 수험자들이 재고자 하는 구인과 무관한 과정을 거쳐서 시험 문제에 답을 하는 경우, 해당 문항은 인지타당도가 낮은 문항이므로, 가능하다면 그 영향을 줄일 수 있는 방향으로 문항을 수정해야 한다는 것이 두 연구의 공통적 주장이다. 면담 이외에 시선 추적(eye-tracking) 방법을 사용해서 읽기 시험(Bax, 2013) 혹은 듣기 시험(Aryadoust, 2020)을 볼 때 수험자들이 거치는 인지 과정을 연구한 연구들 역시 공통적으로 평가의 인지타당도가 평가를 통해 수집되는 정보의 타당도에 중요한 영향을 미친다는 점을 전제로 하고 있다. 최근에는 기능적 근적외 분광법(functional near-infrared spectroscopy; fNIRS)과 같은 신기술을 적용해서 학생이 평가 과업 수행 시 거치는 인지 과정을 밝히는 연구(Aryadoust, Foo & Ng, 2022)도 이루어지고 있는데, 이는 평가의 인지타당도 확인이 평가의 타당도를 논의함에 있어 기본 자료가 된다는 점을 보여준다.

인지타당도의 개념을 2.1. 절에서 논의된 '학생들이 이미 읽어본 글을 지문으로 제시한 읽기 시험'에 적용해보면, 우리나라에서 시행되는 영어 읽기 시험 중 적어도 일부는 진정한 의미의 읽기 시험이 아닐 수 있다는 해석이 가능해진다. 영어 읽기 시험에서 재고자 하는 능력은 영어로 쓰여진 글을 '읽고' 이해할 수 있는 능력이라는 점에는 이견이 없을 것이다. 그러나 읽기 시험에 교과서 혹은 EBS 교재와 같은 학생들이 이미 읽어본 지문을 활용하는 경우, 학생들은 주어진 지문을 '읽지 않고도' 문제에 대해서 답할 수 있다. 시험 상황이 아닌 실제 언어사용 상황에서의 읽기 과정은 주어진 글을 읽는 인지적 과정을 필수적으로 거치게 된다. 따라서 이 과정을 거치게 하지 못하는 문항, 즉 지문을 읽지 않고도 맞게 답할 수 있는 문항은 위에서 설명한 인지타당도가 낮은 문항이며 따라서 읽기 능력에 대한 유의미한 정보를 제공하기가 어렵다. 겉보기에는 읽기 지문과 이에 대한 문제로 구성된 읽기 시험이지만, '읽기' 과정이나 '읽기' 능력에 대해서 제공하는 정보는 상당히 제한적인 것이다.

실제로 Kahng(1995)의 연구에서, 주어진 읽기 지문을 읽어야 독해 문제에 올바르게 답할 수 있는 정도를 단락의존도(passage dependency index)라는 개념을 이용해 연구한 결과, 학생들이 이미 학습한 지문으로 읽기 평가를 구성한 경우에는 지문 없이 문제만 보고 풀었을 때의 정답률과 지문과 문제가 함께 제시되었을 때의 정답률 간의 차이가 크지 않다는 결과를 얻었다. 이러한 결과를 토대로 Kahng(1995)은 학생들이 읽은 적 있는 교과서 지문을 이용해서 읽기 능력을 평가하는 것이 적절하지 않다고 지적하였으며, "독해력 측정도구를 만들 때 기습한 교과서의 본문은 반드시 그 대상에서 제외시켜야 한다"(p. 33)고 주장하였다. 더 나아가 그는 이런 지문을 이용하여 읽기 평가를 실시하는 것은 "독해력을 측정한다고 말하기가 어려울 것이며 지문 단락의 존재가치에도 의문이 갈 것이다"(p. 33)라고 강조한다. 비록 저자가 인지타당도라는 용어를 사용하고 있지는 않지만, 학생들이 지문을 읽지 않는다면 읽기를 제대로 평가할 수 없다는 Kahng(1995)의 주장은 인지타당도와 밀접한 관련이 있는 것으로 해석할 수 있다. 하지만 우리나라 영어 교육에서는 이미 25 년 전에

발표된 이 논문의 제안이 수용되지 못하고, 고등학교 내신평가에서뿐 아니라 국가 수준의 평가인 수능의 읽기 시험에서도 학생들이 이미 읽어본 글이 읽기 평가의 지문으로 사용되고 있다는 것을 앞서 살펴본 Sung 과 Jo(2015), Lee 와 Lee(2017), Kwon(2015)의 논문에서 확인할 수 있다. 수능의 경우, EBS 교재의 연계 비율은 한 때 70% 수준에 달했으나(Lee, 2020), 2022 학년도 수능에서는 약 50%로 낮아졌다(Kim, 2021). 그럼에도 여전히 적지 않은 비율의 수능 읽기 문항들이 학생들이 이미 학습한 지문을 바탕으로 한다는 것을 알 수 있다.

이처럼 우리나라의 영어 읽기 평가가 평가 이론에 부합하지 않는 부분이 있음을 확인해볼 수 있다. 이 사실은 좋은 평가가 갖춰야 하는 특성 중 가장 중요한 요소라고 할 수 있는 타당도(validity)에 대한 논의가 충분히 이루어지지 않은 채 영어 읽기 평가가 실시되고 있음을 보여주는 증거로 해석될 수 있다. 실제로 우리나라 영어 교사들의 언어 평가 전문성(language assessment literacy)를 살펴본 Chung 과 Nam(2018)의 연구에서 참여 교사들은 영어 평가와 관련된 보다 많은 교육과 훈련이 필요하다는 의견을 제시하였으며, 전체 참여 교사의 92%가 타당도와 관련된 전문성을 기를 필요가 있다고 응답한 것으로 나타났다. 연구자들은 또한 교사들이 평가 이론에 대한 교육뿐 아니라 평가 이론을 평가 문항 개발에 적용할 수 있는 실질적인 훈련의 필요성을 언급했다고 보고하고 있다. 다음 장에서는 이와 같은 영어 교사들의 요구를 충족시켜 줄 수 있는 방안 중 하나로 증거기반평가모델(Evidence-centered assessment design; ECD; Mislevy et al., 2003)을 소개할 것이다.

3. 증거기반평가모델과 읽기 평가

이 장에서는 증거기반평가모델을 소개하고, 이 모델의 기본 구성요소에 대하여 간략히 설명한 후, 이 모델이 평가 개발 과정에 적용된 구체적인 예를 제시할 것이다. 이를 바탕으로, 증거기반평가모델을 의식적으로 적용하는 과정에서 평가 출제자나 교사가 평가로부터 수집되는 정보의 질에 대해 본질적인 논의를 할 수 있다는 점을 읽기 평가의 예시를 통해 논의하고자 한다.

3.1. 증거기반평가모델

증거기반평가모델은 "증거 논증의 관점에서 교육 를 구성하기 위한 접근법(an approach to constructing educational assessments in terms of evidentiary arguments)"이다(Mislevy et al., 2003, p. i). 이 모델은 Messick(1994)이 좋은 평가의 근간으로 제시한 '증거추론(evidentiary reasoning)' 이론을 평가의 제작, 구성 및 결과 해석에 적용할 수 있게 하는 구체적이고 실질적인 틀을 제공한다. Mislevy 등(2003)이 설명하는 증거기반평가모델은 다음의 세 가지 전제에 기초하고 있다. 첫째로, 평가에서 재고자 하는 능력, 즉 구인(construct)이 명확하게 정의되어야 한다. 둘째로, 한 학생의 평가 수행 결과를

바탕으로 이 학생이 비평가 상황에서 어떻게 행동할 것인가에 대한 추론이 이루어지며, 평가 과업의 개발에서 결과의 해석에 이르는 일련의 추론 과정은 증거추론의 원칙에 따른다. 그리고 평가의 목적(purpose)이 평가 구성의 핵심 고려사항이라는 점이 세 번째 전제이다(Mislevy et al., 2003, p. 20).

특히 두 번째 전제는 평가의 계획 단계에서 시작하여 제작, 실시, 결과 해석과 활용까지 이르는 일련의 과정을 서로 유기적으로 연결된 연쇄적 추론 과정(chain of reasoning)으로 이해함에 있어서 핵심이 된다. Mislevy 등(2003)이 제안한 증거기반평가모델의 가장 큰 장점은 이 연쇄적 추론 과정을 몇 개의 세부 모델로 구분한 뒤, 각 모델에서 고려해야 하는 사항들을 설명함으로써, 추론 과정이라는 다소 추상적인 개념을 보다 실제적으로 이해하고 적용할 수 있도록 해 준다는 것이다. 그림 1 은 증거기반 평가 모델이 6 개의 세부 모델, 즉 학생모델(student model), 증거모델(evidence model), 과업모델(task model), 조립모델(assembly model), 제시모델(presentation model)과 이 5 개의 모델들을 아우르는 전달모델(delivery model)로 구성됨을 보여준다.

FIGURE 1

Components of Evidence-Centered Assessment Design (Mislevy et al., 2013, p. 5)

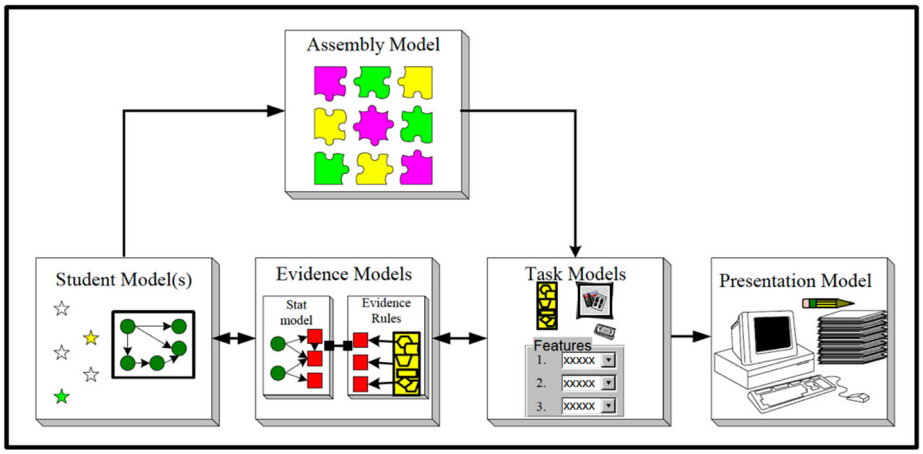


그림 1 에 제시된 6 개의 세부 모델들은 (1) 학생, 증거, 과업모델과 (2) 조립, 제시모델, 그리고 (3) 전달모델의 세 그룹으로 나눠 볼 수 있다. 첫 번째 그룹인 학생, 증거, 과업모델은 개별 평가 문항을 출제할 때 고려되어야 하는 모델들이며, 두 번째 그룹인 조립, 제시모델은 개별 문항들을 조합하여 전체 시험을 구성·제시할 때 고려해야 할 사항들을 다루고 있는 모델이다. 마지막으로 그림의 바깥 쪽에 표시된 전달모델은 위에서 설명한 5 개 모델을 아우르는 모델로서 평가의 플랫폼,

보안(security) 문제, 시간 등 시험 시행에 관한 제반사항을 다루는 모델이다. Mislevy 등(2003)은 평가를 이와 같은 하위 모델로 구분해서 개발하는 것이 평가의 목적에 따라 각 모델들을 다른 방식으로 조합하는 것을 가능하게 한다고 설명한다.

시험의 구성 요소인 개별 문항들을 잘 만드는 것은 전체 시험의 타당도 확보를 위한 필요조건이 된다. 따라서 그림 1 의 학생, 증거, 과업모델의 고려사항에 따라 좋은 문항을 개발하는 것은, 이후 단계에서 개별 문항을 조합하여 한 세트의 시험을 구성하고 전달하는 조립, 제시모델 및 전달모델 단계에서 평가 결과의 타당도를 확보하기 위한 바탕이 된다. 따라서, 아래의 논의에서는 개별 문항 작성에 초점을 두고, 이 때 고려해야 할 사항들을 학생, 증거, 과업모델의 각 단계별로 논의해 보고자 한다. 각 단계의 핵심 고려사항, 즉 각 단계에서 출제자가 자문해야 하는 핵심 질문은 아래와 같다.

1) 학생모델(Student Model)의 핵심 질문은 "어떤 능력을 재고자 하는가(What are we measuring)"(Mislevy et al., 2003, p. 6)이다. 이 질문에 대한 대답은 평가 결과를 바탕으로 우리가 수험자의 능력(ability)이나 지식(knowledge)에 대하여 내리는 해석과 연결된다. 즉, 학생모델은 한 평가에서 재고자 하는 능력, 구인(construct)을 정의하는 단계로 이해될 수 있다.

2) 증거모델(Evidence Model)에서 중심이 되는 질문은 학생모델에서 정의한 평가 구인을 "어떻게 잴 것인가(How do we measure it)"(Mislevy et al., 2003, p. 8)이다. 이 단계는 두 부분으로 나뉘는데, 첫 번째는 평가를 통해 수집한 학생 수행 결과물(work product)의 중요 특징을 판별하고 그 특징을 규정하는 '증거 규칙(Evidence rules)'이며, 다른 하나는 평가 결과에서 나타나는 수행 특성을 채점하여 그 결과를 학생 모델과 연결 짓는 '통계 모델(Statistical model)'이다.

3) 과업모델(Task Model)에서는 "어디서 잴 것인가(Where do we measure it)" (Mislevy et al., 2003, p. 8)가 핵심 질문이다. 이 질문에서 '어디서'는 학생들에게 '어떤 조건에서 평가를 수행하게 할 것이냐'로 재해석될 수 있다. 즉 과업모델의 핵심 질문에 대한 답을 찾아가는 과정에서 어떤 언어 자료를 제시하고, 어떤 과업을 통해서 학생들의 평가 수행 자료를 수집할 것인지를 결정하게 된다.

위와 같이 정의되는 학생, 증거, 과업모델은 그림 1 에서 보듯이 서로 밀접하게 연관되어 있으며, 이 모델들을 서로 연결해 가는 과정이 바로 증거추론(evidentiary reasoning)이라는 논리적 추론 과정이다. 즉, '어떤 능력 혹은 지식에 대한 정보를 필요로 하는가'라는 학생모델의 핵심 질문은 '어떤 증거들이 우리가 관심을 갖는 능력 혹은 지식의 정도를 보여줄 수 있는가'라는 증거모델의 질문으로 자연스럽게

연결된다. 이 증거모델의 질문은 다시 '어떤 유형의 문항 혹은 과업을 사용했을 때 우리가 원하는 증거를 수집할 수 있는가'라는 과업모델의 주요 질문으로 이어지게 된다.

타당도(validity)는 좋은 평가가 갖춰야할 가장 근본적인 특성이지만, 개념의 추상성으로 인해 평가 개발자에게 이해하기 어려운 개념으로 이해될 수 있다. 증거기반평가모델의 가장 큰 장점은 평가 개발 과정을 서로 유기적으로 연결된 하위 모델로 세분하고, 각 모델에서 어떤 구체적인 결정을 내려야 하며 이 결정이 평가의 타당도와 어떻게 관련되는지 명시적으로 확인할 수 있는 틀을 제공(Zieky, 2014)함으로써 평가의 타당도를 높일 수 있는 보다 실천적인 방안을 제시한다는 점이다. 평가 개발자나 교사는 위에서 제시한 각 하위 모델의 핵심 질문을 의식적으로 자문하고 그 자문 결과를 평가 개발에 반영함으로써 평가의 타당도를 높일 수 있는 방향으로 사고할 수 있게 된다. 이러한 이유로 다양한 평가 개발 프로젝트에 증거기반평가모델이 적용되었으며, 특히 Yin 과 Mislevy(2021)는 이 모델을 적용하여 언어 평가를 개발함으로써, 평가하고자 하는 능력에 대한 더 타당한 평가를 개발할 수 있다고 강조한다. 평가 개발에 증거기반평가모델을 적용한 실례로는 TOEFL iBT (Chapelle, Enright, & Jamieson, 2008), 토익 스피킹(TOEIC Speaking) 및 토익 라이팅(TOEIC Writing) 시험(Hines, 2010), 영어 모국어 화자의 논증(argumentation) 능력을 재기 위한 평가(Deane 과 Song, 2014) 등을 꼽아볼 수 있다. 그러나 증거기반모델이 국내의 영어 평가에 적용된 사례는 아직 찾아볼 수 없다.

다음 절에서는 증거기반평가모델의 학생, 증거, 과업모델 각 단계에서 고려해야 하는 핵심 질문이 2 장에서 논의한 영어 읽기 평가에 어떻게 적용되는지 상세하게 기술하고자 한다. 이를 통해 증거기반평가모델의 하위 모델 간 연결성 중 어느 한 부분이라도 어긋나는 경우에는 평가 결과가 타당한 결론에 이르지 못할 수 있다는 것을 확인할 수 있다.

3.2. 증거기반평가모델의 적용: 읽기 평가

그림 1 에 제시된 증거기반평가모델의 하위 모델 중 학생, 증거, 과업모델을 읽기 평가에 적용해 보면 그림 2 와 같이 나타낼 수 있다. 먼저 학생모델에서는 평가에서 재고자 하는 능력, 즉 구인(construct)을 정의하게 된다. 이 능력은 '영어 읽기 능력' 과 같이 일반적으로 기술될 수도 있고 '설명문을 읽고 세부내용 파악하기' 등과 같이 좀더 상세하게 기술될 수도 있다. 학생모델에서 재고자 하는 능력을 규정한 후에는 평가에 포함될 과업의 특성과 그 과업 수행의 성공 정도를 채점하는 방법을 각각 과업모델과 증거모델에서 결정하게 된다.

FIGURE 2

Application of the ECD Model to Reading Comprehension Assessment

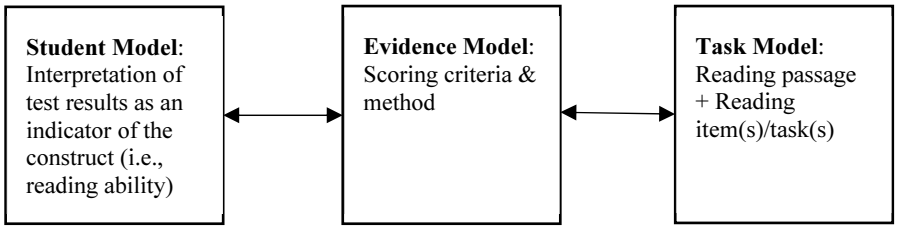
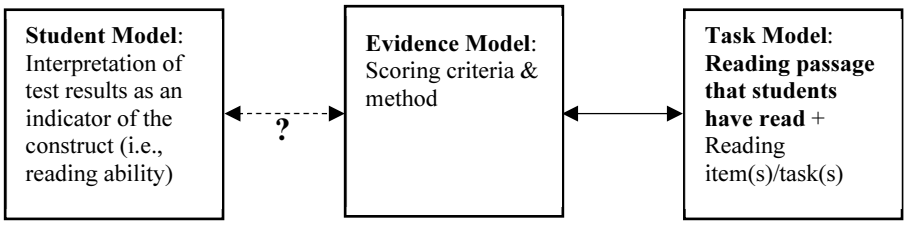


그림 2 에 제시된 3 개 모델 중 가장 구체적인 형태를 띠고 직관적으로 이해 가능한 모델은 과업모델이다. 읽기 지문과 그 지문의 이해를 바탕으로 답해야 하는 문항은 '읽기' 시험의 외형적인 형태를 결정하는 가장 기본적인 구성 요소이며, 이 요소들의 특성을 정하는 단계가 과업모델이다. 과업모델이 이처럼 직관적으로 이해가능한 반면, 학생모델은 평가 결과(예. 정답 여부)가 갖는 의미 해석과 관련된 단계로 이 단계의 논의는 다소 추상적이다. 눈에 보이는 형체가 없기 때문이다. 이러한 이유로 읽기 평가 출제 단계에서는 평가의 구체적인 구성 요소인 읽기 지문의 선정과 각 지문에 적합한 문항 출제에는 많은 시간과 노력을 들이지만, 그런 문항을 통해서 학생들로부터 수집한 정보의 의미 해석에 대해서는 의식적으로 신경 쓰지 않으면 간과하는 일이 발생할 수 있다. 학생모델에 대한 충분한 고려없이, 과업, 증거모델에 집중한 시험 문항의 대표적인 예가 앞서 2 절에서 논의한 학생들이 이미 읽어본 지문을 사용해서 읽기 평가를 구성하는 경우이다. 그 이유를 시각적으로 나타내면 그림 3 과 같다. 그림 2 와 비교해서, 그림 3 에서 주목해야 할 부분은 과업모델에서 빨간색으로 표시된 부분, 즉 학생들이 이미 읽어본 글을 읽기 지문으로 제시한다는 부분이다. 또한 그림 3 에서는 학생모델과 증거모델을 연결하는 화살표가 점선으로 표시되고, 그 아래 물음표가 제시되어 있다. 이는 읽기 지문이 학생들이 이미 읽어본 글인 경우에는 시험 결과를 학생의 읽기 능력을 나타내는 지표로 해석하기 어려울 수 있다는 점을 상징적으로 나타낸다.

FIGURE 3

Application of the ECD Model to Reading Assessment with a Reading Passage Known to Students



읽기 지문으로 어떤 글이 사용되든, 다시 말해 학생들이 이미 읽어본 글을 활용하든 혹은 처음 보는 글을 사용하든, 시험의 외형인 과업모델에는 차이가 없다. 그리고 시험 문항에 대한 응답을 타당한 기준에 따라 채점하기만 한다면 증거모델 단계에서도 특별히 차이가 나는 부분은 없다. 그러나 증거모델에서 학생모델로 넘어가는 단계에서는 어떤 글이 지문으로 사용되었느냐에 따라 상당히 큰 차이가 발생하게 된다. 학생모델에서 고려해야 하는 질문은 '시험을 통해서 학생의 읽기 능력에 대한 유의미한 정보를 얻었는가' 혹은 '시험 점수로 미루어 이 학생의 읽기 능력은 높은가 (혹은 낮은가)' 등인데, 학생들이 이미 읽었던 지문을 사용한 시험의 결과로는 이 질문들에 대한 답을 하기가 어렵다. 그 이유는 학생들이 이미 내용을 알고 있는 글이 읽기 지문으로 제시될 경우, 학생들이 지문을 읽을 필요 없이 기억하고 있는 내용을 바탕으로 평가 문항에 답할 가능성이 높아지기 때문이다. 이는 이미 2.1 절에서 인용된 Lee 와 Lee (2017)의 학생 인터뷰 결과에서도 확인된 현상이다.

지금까지의 논의를 정리해보면, 읽기 시험에 학생들이 이미 읽어봤거나 공부한 지문을 사용하는 경우 학생들은 주어진 지문을 '읽지' 않고, 즉 '읽기'라는 인지적 과정을 거치지 않고 정답을 맞출 수 있게 된다. 이러한 문항은 인지타당도가 낮은 문항이라는 문제가 있으며, 이 문제점은 눈에 보이지 않는 현상이라는 점에서 인식되지 않고 간과되기 쉽다. 아마도 이런 이유때문에 수능에서 EBS 지문을 그대로 사용했던 직접 연계 정책의 근본적인 문제점, 즉 낮은 타당도에 대한 학술적인 문제제기가 없었다[1] 고 추측해볼 수 있다. 그러나 앞서 그림 2와 그림 3의 대조를 통해 논의했듯이, Mislevy 등(2003)의 증거기반평가모델을 적용한다면 학생모델과 같이 구체적 형태로 드러내지 않으나 평가의 타당도 확보를 위해서 꼭 필요한 고려사항을 놓치지 않게끔 해줄 수 있다.

4. 결론 및 교육적 함의

본 논문은 한국 영어교육학계에서 주로 출판되는 논문들과 같이 실증적 일차 자료 분석을 기반으로 한 논문은 아니다. 그러나 이 논문을 통해, 그동안 너무나 기본적인 것임에도 불구하고 계속해서 간과되어 왔던 영어 읽기 평가의 문제점을 직접적으로 제기하고, 학생이 이미 읽었던 글을 이용해 읽기 능력을 평가하는 것이 문제가 되는 이유를 증거기반평가모델을 사용하여 잘 포착해 낼 수 있다는 점을 보여주고자 하였다.

[1] 수능-EBS 연계 정책이 수능 읽기 지문의 난이도에 미친 영향을 살펴보는 연구는 다수 보고되었다(예., Kim & Lee, 2017; Kim & Choi, 2015; Yang & Lee, 2019 등). 그러나 이 정책으로 인해 '수능 영어 읽기 평가 결과를 영어 읽기 능력을 보여주는 지표로 해석하기 어렵다'는 보다 근본적인 문제제기를 한 학술논문은 한국학술지인용색인(KCI) 등재지 학술지에서 찾을 수 없었다.

논문에서는 먼저 학생들이 이미 읽어 본 글을 읽기 지문으로 제시하는 평가 과업으로 학생들의 읽기 능력을 제대로 평가할 수 없는 이유를 평가 이론, 그 중에서도 특히 인지타당도의 개념을 통해 지적하였다. 이미 읽어 본 글을 시험에서 접하는 경우 수험자는 글을 읽을 필요도 없이 읽기 평가 문항에 답을 할 수 있으며, 따라서 이때 수험자가 거치는 인지과정은 처음 보는 글을 읽을 때 거치는 인지과정과 다를 수밖에 없다. 이 경우 해당 읽기 평가의 인지타당도는 낮다고 볼 수 있으며, 이런 문항을 사용한 평가 결과를 수험자의 읽기 능력을 보여주는 지표로 해석하는 것은 정당화되기 어렵다. 이러한 문제점에도 불구하고, 우리나라 고등학교의 내신평가와 수능에서는 학생들이 영어 교과서 혹은 EBS 수능 교재에서 이미 읽어본 글이 읽기 시험의 지문으로 제시되고 있음을 선행연구를 통해 확인할 수 있다(Kwon, 2015; Lee & Lee, 2017; Sung & Jo, 2015).

Messick(1989, 1994)에 따르면 평가의 타당화(validation)는 평가의 결과를 의도한 용도에 사용하는 것의 정당성을 입증하는 논증 과정(argumentation process)이다. 이 타당화 논증 과정은 평가 과업에서 시작해서, 과업을 통해 수집된 응답, 응답의 질에 대한 채점, 채점 결과의 해석 단계를 거쳐 그 해석을 바탕으로 내리는 개인의 능력에 대한 판단에 이르기까지 각 단계를 연결해주는 연쇄 추론 과정으로 이해될 수 있다. 이 일련의 과정 중 어느 한 부분에서라도 논리적 결함이 발생한다면 이는 전체 타당화 과정의 논증의 설득력을 약화시키는 결과로 이어지게 된다. 따라서 낮은 인지타당도는 평가의 타당화 논증 과정 전반의 설득력에 부정적인 영향을 미치게 된다.

본 논문에서 소개한 증거기반평가모델은 평가 개발의 전 과정에서 연쇄 추론 과정으로서의 평가 타당화 논증 과정을 의식적으로 적용해볼 수 있는 실용적인 틀을 제시해 준다는 점에서 그 의의를 찾을 수 있다. Zieky(2014)의 설명과 같이, 증거기반평가모델 평가 개발과 그 결과의 해석 과정에서 고려되는 다양한 요소들이 평가의 타당도에 미치는 영향을 명시화(explicit)해주고 있는 것이다. 다시 말해서, 증거기반평가모델의 모듈화된 구조와 각 하위모델에서 고려해야 하는 핵심질문은 평가 개발자가 평가의 각 요소에 대한 결정을 내릴 때 유의해야 할 점에 대한 명확한 가이드라인을 제시해 준다. 그리고 각 하위모델 간 관계를 인식할 수 있도록 유도함으로써, 평가 개발자는 각 단계에서 내리는 결정이 평가의 타당도에 미치는 영향을 줄곧 인식하게 된다. 본 논문의 3 장에서는 영어 읽기 평가에 증거기반평가모델을 적용해 봄으로써 이 모델이 평가의 타당도 논증 과정에 활용되는 실례를 제시하였다.

Mislevy 와 Haertel(2006)은 증거기반평가모델이 평가 개발 과정에서 이미 적용해오고 있던 것들을 표현만 바꿔서 제시하고 있는 것이라는 비판이 있을 수 있음을 지적한다("Is evidence-centered design just a bunch of new words for things we are already doing? There is a case to be made that it is," Mislevy & Haertel, 2006, p. 17). 이러한 비판에 대해 저자들은 증거기반평가모델의 강점은 평가 개발 과정을 여러 하위모델로 구분해서 살펴보면 겉으로는 상당히 달라 보이는 평가들이

본질적으로는 유사한 문제에 대한 답을 찾아가는 과정이라는 점, 즉 '평가'의 본질을 인식할 수 있게 해주는 데 있다고 주장한다("It [Evidence-centered design] helps us understand what we are doing at a more fundamental level," Mislevy & Haertel, 2006, p. 18). 따라서 증거기반평가모델은 지필평가뿐 아니라 수행평가와 같이 지필평가보다 고차원적인 구인을 평가하는 데 있어서도 적용될 수 있다. 앞에서도 예로 언급했던 TOEFL iBT(Chapelle et al., 2008), TOEIC Speaking 과 Writing(Hines, 2010)과 같은 제 2 언어 능력 평가뿐 아니라, 영어 모국어 화자의 논증(argumentation) 능력 평가(Deane 과 Song, 2014), 더 나아가 간호사의 임상 판단 능력(nursing clinical judgment)과 같은 상위인지 구인에 대한 수행기반 평가 개발(Dickison et al., 2016)에도 증거기반평가모델을 적용한 사례가 있다. 그러나 이 증거기반모델이 한국의 영어 평가 분야에 소개되거나 평가 개발에 적용된 예는 아직 까지 찾아볼 수 없다.

본 논문에서 제시한 문제의식과 증거기반모델의 활용에 관한 제안은 영어 읽기 평가를 현행보다 '제대로' 실시할 수 있는 방안에 대한 더 건설적인 논의를 이끌어낼 수 있을 것이다. 본 논문이 이끌어낼 수 있는 추후 논의의 예로 (1) '성취평가'인 내신평가에서 교과서 '밖' 지문을 사용하는 것의 적절성에 대한 논쟁과 (2) 증거기반평가모델을 활용해 영어 교사의 '언어 평가 전문성(language assessment literacy)'을 신장시키는 방안을 제시해 볼 수 있다.

먼저 중고등학교의 내신평가에서 교과서 지문 외의 글을 사용해서 읽기 시험을 구성하는 것에 대해서는 다음과 같은 반대 의견이 있을 수 있다. 내신평가는 '성취평가(achievement test)'이며, 성취평가는 '수업 시간에 배운 내용을 얼마나 잘 학습'했는지 평가하는 것으로, 이러한 성취평가에 배우지 않은 내용을 포함하는 것은 적절하지 않다는 것이다. 이 반론에 대해서, 성취평가를 구성할 때는 '수업 내용(course content)' 자체보다는 '학습목표(course objectives)'를 기반으로 하는 것이 낫다(Hughes, 2003, p. 15)는 점을 주목할 필요가 있다. 영어 교과서의 읽기 지문도 결국 교육과정에 명시된 성취기준(예. [9 영 03-09] "일상생활이나 친숙한 일반적 주제의 글을 읽고 일이나 사건의 원인과 결과를 추론할 수 있다", Ministry of Education, 2015, p. 36)을 학생들이 달성할 수 있도록 도와주는 도구라는 점을 잊어서는 안된다. 학생들이 특정 교과서 지문'만' 읽을 수 있는 능력을 길러주는 것이 아니라 그 지문을 학습한 결과로 비슷한 난이도의 다른 지문도 읽고 이해할 수 있는 능력을 길러주는 것이 학습목표가 되어야 한다. 이러한 학습목표를 학생들이 얼마나 달성했는지 평가하는 내신평가에서는 이전에 이미 읽어서 굳이 읽을 필요도 없이 풀 수 있거나 혹은 완전히 기억해 내지는 못하지만 일부라도 기억해서 풀 수 있는 문제가 아니라, 글을 실제로 '읽고' 답할 수 있는 문항으로 평가가 구성되어야 한다. 이런 조건이 충족되었을 때 학생의 읽기 성취기준의 달성 여부에 대한 유의미한 정보 수집이 가능하고, 이 정보는 학생의 읽기 능력에 대한 올바른 추론으로 이어지게 된다.

내신평가에서 교과서 지문 이외의 글이 지문으로 제시되는 것이 궁극적으로는 영어 교수·학습에 긍정적인 환류효과(washback)를 가져올 수 있다(McKinley &

Thompson, 2018). 평가의 환류효과는 "교사와 학생에게 무엇을 배워야 하고, 어떻게 가르쳐야 하고, 학생들이 학습 목표를 달성하기 위해서 무엇을 해야 하는지에 대한 분명한 메시지(a clear message … about what should be learned, how it should be taught, and what teachers and students can do to reach their respective goals)"(Chung & Nam, 2018, p. 44)를 전달해야 한다는 맥락에서 이해해 볼 수 있다. 내신평가의 읽기 평가에서 제시되는 지문을 바꾸는 것은 우리나라 학생들이 영어 읽기를 공부하는 데 있어 근본적인 변화를 가져올 수 있다. 학생들은 교과서에 실린 글을 (거의 외울 수 있는 정도까지) 반복해서 읽는 대신 다양한 영어 글을 읽는 방식으로 영어 읽기 공부 방식을 바꾸게 될 것이며, 이것은 학생들의 영어 읽기 의사소통 능력을 향상하는 데 보다 더 긍정적인 영향을 미칠 것이다.

마지막으로 증거기반평가모델이 교사들의 언어 평가 전문성, 평가 문해력(language assessment literacy) 신장을 위한 유용한 도구로 활용될 수 있다는 점을 강조하면서 본 논문을 마무리하고자 한다. Mislevy 와 Haertel(2006), Zieky(2014)에서 논의되었듯이 증거기반평가모델은 좋은 평가를 개발하기 위해서 일반적으로 고려되는 요소들을 이해하기 쉽고, 보다 체계적인 방법으로 제시한 실용적인 틀이다. 그리고, 앞서도 언급했듯이 이 틀은 본 논문에서 예시로 제시한 읽기 평가 이외에 다양한 능력, 지식을 평가하는데 적용될 수 있다는 장점을 갖고 있다. 따라서 이 모델을 예비 교사 양성 과정이나 교사 연수 등 재교육 과정에서 소개하고 실제 평가 개발이나 기존 평가 분석에 적용해 보게 한다면, 교사의 평가전문성 향상에 기여할 수 있을 것이다. 이는 학교에서 실시되는 영어 평가의 질 향상으로 이어질 것이고, 궁극적으로는 긍정적인 환류효과(washback effect)를 가져올 수 있을 것이다.

Applicable level: Secondary

REFERENCES

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

Aryadoust, V. (2020). Dynamics of item reading and answer changing in two hearings in a computerized while-listening performance test: An eye-tracking study. *Computer Assisted Language Learning, 33*(5-6), 510-537.

Aryadoust, V., Foo, S., & Ng, L. Y. (2022). What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessment*? Language Testing, 39*(1), 56-89.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing, 30*(4), 441-465.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1-26). New York, NY: Routledge.

Cho, Y., & So, Y. (2014). *Construct-irrelevant factors influencing young English as a foreign language (EFL) learners' perceptions of test task difficulty* (Research Memorandum No. RM-14-04). Princeton, NJ: Educational Testing Service.

Chung, S. J., & Nam, Y. (2018). Language assessment literacy of Korean EFL teachers: An investigation of their training experiences and needs. *Modern English Education, 19*(1), 38-48.

Deane, P., & Song, Y. (2014). A case study in principled assessment design: Designing assessments to measure and support the development of argumentative reading and writing skills. *Psicología Educativa, 20*, 99-108.

Dickison, P., Luo, X., Kim, D., Woo, A., Muntean, W., & Bergstrom, B. (2016). Assessing higher-order cognitive constructs by using an information-processing framework. *Journal of Applied Testing Technology, 17*(1), 1-19.

Field, J. (2013) Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (Studies in Language Testing Vol. 35, pp. 77-151). Cambridge, England: UCLES/Cambridge University Press.

Hines, S. (2010). *Evidence-centered design: The TOEIC Speaking and Writing tests* (TOEIC Compendium 7.12). Princeton, NJ: Educational Testing Service.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.

Kahng, Y.-K. (1995). Passage dependency in high school English reading tests. *English Teaching, 50*(1), 21-36.

Kim, J., & Lee, D. J. (2017). A corpus-based study on the use of vocabulary in high school English I·II textbooks, College Scholastic Ability Tests, and EBS materials. *The Journal of Linguistic Science, 80*, 51-74.

Kim, J. E., & Choi, I.-C. (2015). A corpus-based comparative analysis of linguistic difficulty among high school English textbooks, EBS-CSAT prep books, and College Scholastic Ability Test. *Multimedia-Assisted Language Learning, 18*(1), 59-92.

Kim, J. J. (2021, November 18). English section of the College Scholastic Ability Test: The linkage rate of EBS textbooks dropped. *Hankookilbo*. https://m.hankookilbo.com/News/Read/A2021111817540003966

Korea Institute for Curriculum andd Evaluation. (2016). *A study guide to the criterion-referenced assessment of the English section of the College Scholastic Aptitude Test*. Korea Institute for Curriculum and Evaluation. https://www.suneung.re.kr/boardCnts/fileDown.do?fileSeq=0f946c8e039659ca355e82e21e850bb1.

Kwon, O. (2015). A history of policies regarding the English section of Korea's College Scholastic Ability Test. *English Teaching, 70*(5), 3-34.

Lee, D. R., & Lee, H. (2017). Exploring high school students' and in-service teachers' perception on difficulty of a regular English reading achievement test: Focus on the item type. *English Language Assessment, 12*, 99-118.

Lee, H. (2020). Vocabulary analysis of CSAT English tests and CSAT-EBS preparation coursebooks, with reference to the reading tests. *Modern English Education, 21*(3), 48- 57.

McKinley, J., & Thompson, G. (2018). Washback effect in teaching English as an international language. In *The TESOL encyclopedia of English language teaching*. https://doi.org/10.1002/9781118784235.eelt0656.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: American Council on Education & Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Ministry of Education. (2015). *2015 National English Curriculum*. https://ceri.knue.ac.kr/pds/2015_02_english.pdf

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief-introduction to Evidence-centered design* (Research Report RR-03-16). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement Issues and Practices, 25*(4), 6-20.

Sung, M., & Jo, V. H. (2015). Sociopolitical and contextual influences on teacher-produced achievement tests of English in Korean high schools. *SNU Journal of Education Research, 24*, 115-148.

Winke, P., Lee, S., Ahn, J. I., Choi, I., Gui, Y., & Yoon, H.-J. (2018). The cognitive validity of child English language tests: What young language learners and their English-speaking peers can reveal. *TESOL Quarterly, 52*(2), 274-303.

Yang, S., & Lee, D. J. (2019). A corpus-based analysis of the topic distribution and vocabulary level of textbooks, EBS materials, and CSATs. *Journal of Learner-Centered Curriculum and Instruction, 19*(4), 711-729.

Yin, C., & Mislevy, R. J. (2021). Evidence-centered design in language testing. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 289-305). London, England: Routledge.

Zieky, M. J. (2014). An introduction to the use of evidence centered design in test development. *Psicología Educativa, 20*, 79-87.