채점 수 누적에 따른 쓰기 평가 과정의 눈동자 움직임 연구

이지원(제1저자) 한국교원대학교 국어교육과 박사 과정
박영민(교신저자) 한국교원대학교 국어교육과 교수

Ⅰ. 서론

학생 글을 평가하는 것은 평가자에게 높은 인지적 부담을 주는 복합적이고 어려운 과제이다(Hamp-Lyons & Henning, 1991). 학생 글을 비판적으로 읽으면서 그 글의 특성을 분석적으로 파악해야 할 뿐만 아니라 평가 기준을 근거로 하여 질적 특성을 서술하거나 질적 특성을 점수로 해석해야 하기 때문이다. 쓰기 평가 과정에서 평가자는 학생 글의 의미나 주제를 파악하기 위해 꼼꼼히 읽어야 하고 평가 기준을 적용하여 이를 구현하는 방식이 적절한지를 따져 보아야 하며 필요에 따라 학생 글을 상대적으로 비교하여 우열을 가리기도 해야 한다. 이처럼 쓰기 평가는 학생 글에 대한 평가적 판단을 위해 고등 사고 기능의 인지적 처리 과정을 거친다는 점에서 문제 해결 활동으로 부를 수 있다(Deremer, 1998: 13).

대부분의 수행 평가에서는 평가자를 돕기 위해 평가 기준, 평가 기준별 척도 및 특성 서술로 구성된 평가 기준표(rubric)를 활용한다. 평가 기준표는 “쓰기의 내용, 조직, 표현과 같이 수행에서 중요한 것 또는 평가 기준을 목록화한 평가 도구로, 각 평가 기준에 대한 등급의 수준을 분절적으로 표현한 것”(Andrade, 1997: 14)으로서 평가자가 학생의 수행에 다각도로 접근할 수 있도록 돕는 평가 도구이다. 평가 기준표를 명확하게 작성하면 평가 기준을 일관성 있게 적용하는 데 도

움을 줌으로써 평가 결과의 신뢰도와 타당도를 높이는 데 기여할 수 있다.

그러므로 쓰기 평가 도구 개발 단계에서 평가 기준표를 명확하게 작성하여 평가 계획에 포함하는 것은 체계적인 쓰기 평가를 위한 필수적인 과정이라 할 수 있다. 그럼에도 불구하고 쓰기 평가 과정에서 평가자들이 이를 어떻게 활용하는가에 대한 정보는 매우 부족하다. 예를 들면, 쓰기 평가자 연수[1]에서 예시문은 몇 편 정도를 채점해야 적절한지, 점수 조정이 필요한 평가자 간의 차이는 얼마인지, 평가자들이 평가 기준표를 적용할 때 어떤 어려움이 있는지 등에 대해 알려진 바가 거의 없다.

이러한 문제를 해소하기 위해 최근 쓰기 평가 연구에서는 평가자 연수 프로그램을 체계적으로 구성하기 위한 논의가 이루어지고 있다. 쓰기 평가의 신뢰도 제고를 위한 평가자 연수는 위에 언급한 문제들을 부분적으로 해소하는 데 도움을 줄 수 있다. 이러한 상황을 고려하여 이 연구에서는 눈동자 움직임을 추적·분석하여 쓰기 평가의 평가자의 평가 과정에 대한 실증적인 데이터를 제공하고자 하는 데 목적이 있다.

1. 평가 기준의 표상 과정

평가자들은 모든 평가 기준들을 기억하지는 않지만, 계속 보지도 않는다. 그렇다면 평가 기준은 어떻게 평가자들에게 인식되고 활용되는가? 평가자 인지 연구는 평가자들의 이에 대한 상세한 정보를 제공해 준다. Freedman & Calfee(1983)은 '평가 기준표의 심적 표상'과 '글에 대한 이미지'를 서로 비교함으로써 점수에 대한 판단을 내리는 평가

1 이 글에서 사용한 '평가자 연수'라는 용어는 여러 쓰기 평가 연구에서 사용해 온 '평가자 훈련'과 차이가 있다. 평가자 훈련은 쓰기 평가의 전문성이 부족한 사람에게 기본적인 소양을 교육하는 프로그램을 뜻한다면, 평가자 연수는 쓰기 평가의 기본적인 소양을 갖춘 평가자들이 신뢰도를 유지하도록 하기 위한 교육이나 협의를 뜻한다(박영민, 2015; 박영목, 2008).

과정의 기초적인 모형을 제안하였다. 평가자들은 채점 과정에서 평가 기준표의 표상화를 이루어나가며, 글에 대한 채점이 축적되어 나갈수록 평가 기준표의 심적 표상이 점차 확립된다. Wolfe(1997)는 평가자 인지 모형을 채점 절차에 대한 심적 스크립트[2]로서의 '채점 체제'와 평가 기준표에 포함된 각 평가 기준의 심적 표상인 '쓰기 체제'의 두 부분으로 나누었다. 즉 평가자들은 평가 기준에 대한 나름의 심적 표상을 구성하여 이를 학생의 글에 적용하면서 채점을 해 나가게 된다. 이는 평가자들이 평가 기준을 적용해나가는 과정에서 평가 기준을 다양한 방식으로 이해하고 그것을 평가에 활용한다는 것을 의미한다.

실재하는 대상에 대한 이러한 심적 표상(representation)을 구성해나가는 과정은 추상적이고 정신적인 것으로 환원되는 과정이다. 다시 말해 실물 자체가 아니라, 다시(re-) 나타냄(presentation)의 결과로 추상화된 결과물인 것이다(이정모 외, 2002). 이는 동일한 평가 기준표를 가지고 평가를 하더라도 평가자 간에 평가 결과의 차이가 발생하는 주요 원인이며, 평가자 연수가 필요한 이유도 이러한 표상이 처리되는 과정에서 일어날 수 있는 다양한 오차들을 조정하도록 돕는 절차인 것이다. 그런데 평가자가 보여주는 오차 중에서도 그 소재가 내적인 것인지, 외적인 것인지에 따라 평가 결과에 미치는 영향은 다를 수 있다. 예를 들어 다른 평가자들에 비해 특정 평가 기준에 대해 일관되게 낮은 점수를 부여한 '평가자 A'와 특정 평가 기준에서 동일한 수행 수준을 가진 글에 어떤 때는 점수를 높게 부여했다가 어떤 때는 점수를 낮게 부여하는 '평가자 B' 중 어떤 평가자가 더 부적격한 평가자인가?

당연히 평가자 간 오차보다도 평가자 내 일관성이 흔들리는 평가자 B가 부적격한 평가자임이 당연하다. 게다가 최근에는 평가자 간 엄격성의 오차 가능성이 모형화되고 수학적인 보완이 가능해지면서(Weigle, 1998: 281), 여전히 평가자 내 일관성의 문제는 평가 결과의

2 정형화된 상황에서 통상 일어나는 일련의 사건들에 관한 도식적 지식을 스크립트(script)라고 한다(Schank & Abelson, 1977). 즉 스크립트는 관에 박힌 사건들에 관한 도식으로서, '극장 가기', '데이트하기', '인터넷하기' 등 일상생활에서 친숙한 상황들에 대해 형성되어 있다(이정모 외, 2002).

신뢰도에 영향을 미치는 가장 큰 문젯거리로 남아있다. 최근 쓰기 평가 연구에서 활발히 사용되고 있는 다국면 Rasch 모형은 적합-부적합 평가자의 기준으로 개별 평가자의 내적 일관성을 기준으로 하고 있다. 게다가 평가자들 간 엄격성의 차이는 모형을 통해 어느 정도 조정이 가능하기 때문에, 사실상 평가자 연수의 궁극적인 목적은 일관성에 의해 부적합으로 판정된 평가자에 대한 재교육에 있다(Lunz, Wright & Linacre, 1990). 평가자 연수가 평가자들을 서로 똑같아지게 만들지는 못하지만, 내적으로 일관되게 할 수는 있다는 Lunz, Wright, & Linacre(1990), Stahl & Lunz(1991) 등의 연구 결과 또한 이를 뒷받침한다.

사실상 엄격성 수준에 따른 평가자 간 개인차는 불가피한 것이며 독립성 또는 평가에 대한 다양한 관점의 범위로 수렴할 수 있다. 한 편의 글에 대해 평가자 집단에서 일치에 이르도록 강요된 과정이 평가자들 고유의 전문성과 판단을 활용하지 못하도록 할 수 있기(Barritt, Stock & Clark, 1986; Huot, 1990) 때문이다. 따라서 평가자 연수는 평가자 간 오차를 줄이기 위한 목적보다도 평가 기준에 대한 내적 일관성을 유지할 수 있도록 하여 부적합한 평가자를 최소화하는 데에 초점을 두어야 한다. Colton et al.(1997)은 평가자의 비일관성이 부적절한 평가자 연수, 부적절한 채점 기준표의 세부 항목, 또는 '평가 기준표를 내면화하는 평가자의 미숙함'에서 기인한다고 하였다. 평가자가 일관성을 유지하기 위해서는 평가 기준에 대한 내면화 과정에서 확고한 정신적 표상을 구축하는 과정이 필요하며, 평가 시행 기간 내내 그것이 일관되게 유지되어야 한다. 평가자 연수에서 글 평가의 반복적인 연습이 필수적인 것도 이 때문이다. White(1984)에 따르면, 평가자 연수 세션의 목적은 평가자들이 예시글(anchor texts)과 더불어 기술어가 혼합된 평가 기준표를 내면화하도록 돕는 것에 있다.

그렇다면 평가자들은 몇 편 정도의 글을 읽음으로써 평가 기준표를 표상화하는가. 평가자들은 평가 활동의 일부로서 평가 기준표에 진술된 내용을 추상화하고, 평가 기준표의 언어를 해석하며, 이러한 해석을

텍스트의 구체적인 부분에서 조정해나가는 과정이 필요하다(Deremer, 1998: 13). 일반적으로 평가자 연수 프로그램에서는 평가자들의 평가 기준의 확고한 표상을 구축하기 위하여 일정 분량의 예시글 채점을 통해 개별 연습 채점 회기를 거치도록 하고 있다. 예를 들어 TOEFL Writing 영역에 대한 온라인 평가자 연수에 대한 안내 사항을 소개하면 다음과 같다.

> 평가자들은 최소한 한 세트의 글을 채점하는 연습을 통해 각 세션을 시작한다. "조정(calibrate)"은 TOEFL iBT 기준에 대한 그들의 판단을 돕는다. 이러한 과정은 공정하고 정확한 채점을 보증한다. 조정이나 실제 채점을 하는 동안 의문이 생기면, 평가자와 평가 조장은 문제를 논의하기 위해 서로 전화를 할 수 있다(ETS, 2007).

평가자가 내적으로 평가 기준에 대한 이른바 영점을 맞추는 과정을 거치게 되는데 이를 조정(calibration)이라 한다. 개별 평가자들은 해당 쓰기 평가에 대한 예시글을 제공받아 연습을 통한 연수를 통해 평가 기준에 대한 표상을 일관되게 구성할 수 있게 된다.

미국의 학업성취도 평가인 NAEP에서는 조정 세트(calibration set)를 각 평가자들에게 제공한다.[3] 평가자 연수에 쓰이는 연습용 예시글은 대략 10편에서 20편 정도이다. 이 세트는 해당 평가의 실제 답안에서 샘플링된 것일 수도 있고 사전에 연수를 위해 작성된 것일 수도 있다. 호주 NAP는 평가 예시문을 활용하여 보다 구체적으로 평가자들에게 예시글에 대한 정보를 제공한다. 각 채점기준별 항목 점수에 대한 예시글을 제시함으로써 평가자의 조정 과정을 돕는다. 각 평가 기준 10개에 대한 7점 척도의 점수별로 예시문을 제시하며 약 20편 내외의 예시글을 제시한다(NAP, 2013). 이러한 대단위 쓰기 평가 상황에서는 평가자의 오차를 줄이기 위한 평가자 연수가 필수적이다. 샘플 채점 과정에서 평가자들은 글을 읽어 나가면서 자신의 평가 기준에 대한 표

3 http://nces.ed.gov/nationsreportcard/tdw/scoring/scoring_calibration.aspx 2015. 5. 13. 검색.

상을 마치 '눈금을 매기듯이(calibrate)' 정밀하게 조정해나가게 된다.

평가자의 평가 기준표에 대한 표상 과정을 돕기 위해 평가 기준을 최대한 상세하게 제시하는 방안도 있다. 평가 기준표에 대한 연구(Moskal & Leydens, 2000)에서는 이러한 평가자의 오차를 줄일 수 있는 방안으로 평가 기준표 상에 제시된 평가 기준과 수행에 대한 기술어를 보다 상세화하는 방안을 제시하였다. 그러나 상세화된 평가 기준은 오히려 길이 효과로 인하여 평가자의 인지적 부담을 가중시킬 수 있다는 점에서 논란의 여지가 있으며, 이를 통제하기 위해 한정 없이 구체화할 수도 없는 노릇이다.

따라서 이 연구에서는 평가자가 쓰기 평가 기준을 확고한 표상에 도달하는 과정에 초점을 두었다. 이 과정을 밝혀낸다면 평가 기준을 어떻게 설정하고 평가자들이 평가 기준을 채점에 적용하는 과정에 도움이 될 수 있는 전략을 추출할 수 있을 것이라 기대하였다. 이에 따라 쓰기 평가자들이 평가 기준과 글을 볼 때 채점 누적 횟수에 따라 어떻게 변화해 나가는지를 관찰하였다.

2. 평가 과정에 대한 눈동자 움직임 추적

쓰기 평가의 평가자 요인과 관련된 연구는 매우 많지만, 평가 결과 데이터가 아닌 평가 과정에 대한 데이터를 통해 평가자 요인에 접근한 연구는 많지 않다. Hamp-Lyon(1990)는 쓰기 능력의 직접 평가에서의 판정과 평가자 신뢰도에 관한 대부분의 기존 연구들이 평가자보다는 신뢰도를 낮추는 주요 원인에 초점을 잘못 설정하고 있다는 점을 지적한 바 있다. 주로 평가자가 지닌 특성, 예를 들면 성별, 경력, 피로도 등의 요인을 평가 결과와 관련지어 평가자 효과를 밝히는 연구들이 평가 과정에 대한 단면들을 제공해왔다면, 평가 과정에 대한 연구는 평가가 어떻게 이루어지는지에 대한 상세한 정보를 제공한다는 점에 가치가 있다.

인간의 인지 과정을 밝히기 위한 연구 방법으로 가장 대표적인 것이

사고구술 프로토콜 분석 방법(Ericsson & Simon, 1980)이다. 사고구술 프로토콜을 활용한 평가자 연구(Cumming, 1990; Vaughan, 1991; Deremer, 1998; Smith, 2000; Cumming et al., 2002; Boyd et al., 2009, Wolfe et al., 1998; Wolfe, 2005; Lumley, 2005; Milanovic, Saville, & Shuhong, 1996; Sakyi, 2000)는 평가 과정에서 평가자들이 대체로 글을 읽고 분석하고 비교하여 판단을 내리는 과정에서 다양한 의사결정전략(decision-making strategy)을 활용하여 글에 대한 점수를 부여하는 과정을 보여주었다.

그러나 문제 해결, 사고와 같은 고차적 인지 활동의 과정을 밝히는 방법은 점차 다양해지고 있다. 19세기 말 내성법에서부터 시작된 인지 과정 연구 방법은 최근 EEG, 눈동자 움직임 추적, fMRI 등 기술의 발전으로 연구의 범위가 보다 확장되었다.

따라서 이 연구에서는 눈동자 움직임 추적을 통해 평가자의 평가 기준 내면화 과정의 탐색을 시도하고자 한다. '눈은 읽기의 구성적 과정을 보여주는 마음의 창'(Paulson & Freeman, 2003: 345)이다. 따라서 눈동자의 움직임을 파악하는 것은 시각적 지각과 이해의 세계를 파악하는 중요한 단서가 된다. 눈동자 움직임은 시각적 데이터에서 산출되는 빈도, 시간, 장소, 패턴 등에 대한 정량화된 정보를 제공하며, 이를 시지각적 정보뿐만 아니라 이해 및 기억과 관련된 인식의 지표로 활용한다. 따라서 이 연구에서는 인간의 이해와 숙달의 과정을 '보는 행위'의 행동률로 파악하였다.

쓰기 평가 과정에서 눈동자 움직임 추적을 활용한 연구는 많지 않다. 읽기에 대한 눈동자 움직임 추적 연구가 읽기 과정의 많은 속성을 밝혔듯이, 쓰기 평가 또한 눈동자 움직임이 보여줄 수 있는 잠재적인 결과가 기대되는 분야이다. 쓰기 평가 역시 본질적으로 학생 글에 대한 읽기와 이해, 판단과 같은 유사한 인지 활동이 복합적으로 일어나는 활동이기 때문이다.

쓰기 평가에서 눈동자 움직임을 활용한 연구로는 박영민(2014)이 있다. 이 연구에서는 국어교사의 평가 과정을 눈동자 움직임 추적을

통해 탐색하였다. 이 연구에서의 변인은 학생 글의 전사 방식이었으며, 워드와 손글씨로 작성된 글에 대한 평가 과정의 차이를 분석하였다. 그 결과 손글씨 설명문이 눈동자 고정 시간과 빈도에서 더 높은 수치를 보였다는 것이 나타났다. 이는 평가자들이 손글씨 설명문을 채점하는 데서 오는 인지적 부담을 반영한다. Winke & Lim(2015)는 ESL 쓰기 평가자의 인지적 처리과정을 분석하여 평가자들이 평가 기준표의 하위 요소에 대한 눈동자 움직임에 대한 차이를 분석하였다. 그들은 평가자들이 주로 평가 기준표 좌측 상단에 위치한 평가 기준들에 주의를 집중하였다는 점에 주목하고 평가 기준표의 시각적 구성 방안에 대하여 논의하였다.

특히 이 연구에서 활용하고자 하는 눈동자 움직임 연구의 장점은 특정 활동의 인지적 처리가 시간적으로 어떻게 전개되는지에 대한 정보를 제공한다는 데 있다. 우리는 눈동자의 움직임을 추적함으로써 평가 과정의 흐름에 따른 지도를 만들 수 있다. 이 연구에서는 눈동자 움직임 데이터를 통해 평가자의 쓰기 평가 과정을 분석하고 채점 수 누적에 따른 변화 양상을 분석하고 예측하였다.

3. 연구 질문

이 연구는 눈동자 움직임 추적을 통해 쓰기 평가자가 채점 수가 누적되어감에 따라 평가 기준을 조정해나가는 과정을 탐색하기 위하여 수행되었다. 쓰기 평가 과정에서 평가 기준은 글과 함께 제시되는 자료로서 평가자의 읽기 대상이며, 따라서 눈동자 움직임 추적을 통해 얻은 데이터는 이 과정에 대한 많은 증거를 제공할 수 있으리라 기대하였다. 이 연구는 평가자의 평가 기준에 대한 확고한 표상이 구축되어가는 과정에서 나타나는 눈동자 움직임의 양상에 초점을 두고 다음과 같은 세 가지 연구 질문을 다룬다.

첫째, 평가자들은 채점 과정 중에 평가 기준표와 글을 보는 빈도와 시간에 어떤 차이가 있는가? 둘째, 평가자들이 평가 기준을 보는 고정

빈도와 시간이 채점 수 누적에 따라 어떻게 변화하는가? 마지막으로, 평가자들은 여러 편의 글을 채점해 나가면서 인지적으로 평가 기준을 구성하게 된다. 이 연구에서는 평가자들이 평가 기준표를 보는 행동의 소거 시점이 평가 기준에 대한 명확한 표상이 형상된 시점으로 판단하고, 평가를 시작한 시간부터 평가를 마칠 때까지의 데이터를 통해 모형을 생성하고, 이 모형의 적용을 통해 채점 수 누적에 따른 채점 과정의 눈동자 움직임 예측값을 추정하였다.

이 연구를 통해 기대되는 연구 성과는 다음과 같다. 첫째, 쓰기 평가자 연수에서 적절한 학생 예시글 편수를 설정할 수 있다. 이 연구를 통해 대부분의 평가 기준 내면화에 평가자들이 소모하는 인지적 자원을 파악한다면 쓰기 평가 평가자 연수 기간에서 사전에 예비 채점에 드는 적절한 시간과 비용을 산출할 수 있다. 둘째, 평가 기준에 대한 인지적 부담 감소 방안을 마련할 수 있다. 평가자들이 평가 기준을 내면화하는 데 있어 어려운 지점이 무엇인지 각 평가 항목에 대한 눈동자 움직임 추적을 통해 파악할 수 있다.

눈동자 움직임 추적 연구는 쓰기 평가의 평가자에 대한 인지 과정을 정량적으로 측정하여 실증적 데이터를 제공하는 데 기여할 것이다. 이 연구는 실제 평가 과정에 대한 연구로서, 쓰기 평가자와 평가 관리자, 평가 기준을 연구하는 연구자들, 그리고 궁극적으로는 학생 모두에게 많은 도움이 될 것이다.

Ⅱ. 연구 방법

1. 참여자

이 연구에는 쓰기 평가 경력 5년 이상의 국어교사 11명이 참여하였으며, 교정 시력이 모두 정상이었다. 남 1명, 여 10명의 국어교사의 평균 연령은 34세(범위 29~42세), 교육경력 10년(범위 5~16년)이었

다. 11명의 평가자 중 1명의 평가자 데이터가 화면 상 좌측 손실률이 높아 분석 대상에서 제외되어 최종적으로 10명의 평가자 데이터를 활용하여 분석하였다.

2. 도구

1) 평가 기준표

이 연구에 사용된 평가 기준표는 Spandel & Culham(1996)의 평가 기준표를 박영민·김승희(2007)가 일부 수정한 평가 기준표로 5점 척도의 '내용, 조직, 표현, 형식 및 어법, 단어 선택'으로 구성되어 있다. 평가 기준표의 선정 과정에는 우리나라에서 일반적으로 박영목(1999)에서 제시한 '내용, 조직, 표현'의 3개 평가 기준을 쓰기 평가에서 주로 활용하고 있다는 점과, 평가 기준표에 대한 최근의 눈동자 움직임 연구(Winke & Lim, 2015)에서 가장 널리 알려지고 활용되는 평가 기준표로 소개된 Jacobs et al.(1981)의 '내용, 조직, 어휘, 언어 사용, 철자법'의 5개 범주가 위에 제시된 Spandel & Culham의 평가 기준과 유사하면서도 보다 한국 실정에 맞다는 점이 고려되었다.

2) 글

고등학교 2학년을 대상으로 수기 작성된 수집된 설명문 30편을 스캔하여, 실제 쓰기 평가 상황과 유사하게 설정하였다. 일반적인 쓰기 평가 상황에서 한 반의 인원수가 30명 정도이고 채점 대상인 점, 그리고 누적 채점에 따른 추세 파악을 위하여 평가자에게 30편의 글을 배정하였다. 쓰기 과제는 '자신이 관심 있거나 좋아하는 책, 영화, 운동(스포츠), 취미 활동 등을 학급 친구들에게 소개(설명)하는 글을 쓰되, 좋아하게 된 이유를 포함하여 쓰시오'였다. 제시되는 글의 순서는 순서 효과의 배제를 위해 두 개의 유형으로 제시되었으며, 동일한 30편의

글을 6명의 평가자에게는 정방향으로, 5명의 평가자에게는 역방향으로 설계된 글 순서로 제시하였다.

3) 장비

눈동자의 움직임은 Tobii glasses eye tracker(Tobii, 스웨덴 스톡홀름)로 측정되었다. 이 장치는 휴대용으로 초당 30번 눈동자의 위치를 파악하여 전송한다. 자료 제시를 위해 해상도 1920×1200 DELL사의 UltraSharp U2410 24인치 모니터 2대를 사용하였다. 모니터 1(좌)에는 PDF 파일로 평가 기준표를 제시하였으며, 모니터 2(우)에는 스캔된 학생 글을 순차적으로 제시하였다. 평가 기준표는 수직주파수가 60Hz로 설정된 DELL 모니터의 피벗 형태로 전환된 세로 모드에서 10포인트 바탕체 서체로 제시되었으며, 화면상으로는 20pixel의 크기로 주사되었다. 참여자들은 양눈으로 글을 읽었으나 오른쪽 눈의 안구운동만이 측정되었다.

3. 연구 절차

실험은 개별적으로 진행되었으며, 참여자들은 실험에 대한 안내를 듣고 휴대용 눈동자 움직임 추적 장치를 착용하였다. 참여자의 정면에 IR Marker를 제시하여 9개의 점을 따라 장치를 조절하였다. 정확도와 적합도가 각 최대 5점인 장치에서 평균 3점 이상인 경우, 유효한 측정치로 판단하여 실험을 진행하였다. 장치 조절 과정을 마친 참여자는 실험용 모니터 앞에 앉아 실험에 대한 안내사항을 읽고, 간단하게 듀얼 모니터와 실험 자료의 조작 방법을 익혔다. 본 실험 시작 전에는 몇 가지 그림 자극을 제시하여, 눈동자의 움직임을 관찰하였으며 채점이 시작되면 참여자들은 제시된 글을 순차적으로 읽어나가면서 각 평가 기준에 대한 점수를 소리 내어 말하도록 하였다. 한 편의 글을 다 읽고 나서 마우스 스크롤을 조절하여 총 30편의 글을 읽었으며, 연구

자가 점수를 받아 적고 이후 녹화된 데이터를 통해 점수를 재확인하는 절차를 거쳤다. 모든 참여자들은 본인의 의사에 따라 채점에 소요되는 시간을 자유롭게 조정할 수 있었으며, 최종적으로 본 실험에 소요된 시간은 평균 30분(범위 19.38~61.35분)이었고, 모든 절차를 포함한 전체 실험은 총 45분 정도가 소요되었다.

4. 데이터 분석

눈동자 움직임 데이터는 Tobii Studio 2.0을 사용하여 분석하였다. 평가자의 30편의 글에 대한 각각의 눈동자 움직임을 분할하였다. 평가자 10명에 대한 전체 글 30편 채점에 대한 유효 데이터 수는 274개[4]였다. 평가 기준표와 글에 대한 눈동자 움직임의 분석을 위해 평가 기준과 글에 대한 AOI(Area of Interest)를 설정하였다. 총 2개의 AOI를 설정하였으며, 평가 기준표에 대한 AOI 1을, 학생 글 전체 영역에 대한 AOI 2를 설정하였다. AOI 설정 예시는 아래의 <그림 1>과 같다.

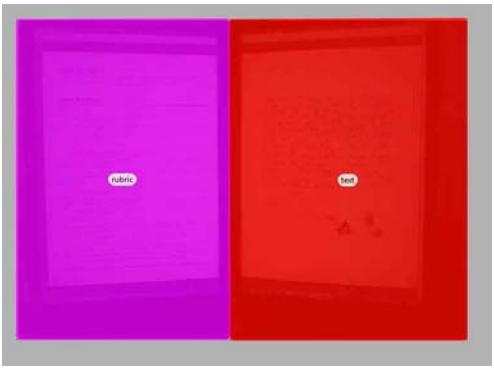

그림 1. AOI 설정 예시(평가자 3)

4 2명의 평가자 데이터가 전반부 17번 글까지 기록되고 18번 데이터부터는 손실되어 8명 평가자의 30편 데이터인 240개 데이터에 2명 평가자의 17편까지의 데이터 34편이 포함되어 총 유효 데이터는 270개로 기록되었다. 손실된 데이터는 결측값으로 처리하였다.

또한 각 평가자의 채점 누적 횟수에 따른 데이터의 측정을 위해, 평가자 채점 자료를 각 글에 대한 채점 횟수별로 분할하고 시계열 데이터로 코딩하였다. SPSS 18.0을 활용하여 기술통계 및 AOI 간 차이 검정, 채점 수 누적에 따른 분석 및 비선형 회귀 모형을 통한 예측값의 추정을 실시하였다.

Ⅲ. 연구 결과 및 논의

1. AOI별 눈동자 움직임의 고정 빈도와 시간

평가자들의 눈동자 움직임의 고정 빈도를 AOI에 따라 구분한 결과, 전체 평가 과정 중 평가 기준표의 고정 빈도로는 약 25%, 고정 시간으로는 약 28%의 비율로 평가 기준표에 눈동자를 고정하였다. 글에 대한 고정 빈도는 약 75%, 고정 시간은 약 72%의 비율로 나타났다. 이 연구에 참여한 평가자들은 학생 글 한 편을 채점할 때 평균적으로 글에는 41.4초, 평가 기준표에는 16.2초의 고정 시간의 분포를 보였다. 평가 기준표와 글에 대한 시선 분포의 차이는 고정 빈도와 고정 시간 모두 통계적으로 유의미하였다.

표 1. AOI에 따른 고정 빈도와 고정 시간

	AOI	N	M	SD	df	t	p
고정 빈도(회)	평가 기준표	274	368.12 (25.06%)	543.37	546	−16.35	.00
	글	274	1100.56 (74.94%)	504.51			
고정 시간(초)	평가 기준표	274	16.20 (28.13%)	22.34	546	−14.08	.00
	글	274	41.40 (71.87%)	19.45			

표 2. AOI에 따른 방문 빈도

AOI	N	M	SD	df	t	p
평가 기준표	274	7.29	7.29	546	−.41	.69
글	274	7.54	7.23			

평가자들의 방문 빈도는 각 관심 영역 내에 고정이 일어난 횟수를 말하는데, 방문 빈도에서는 AOI에 따른 차이가 나타나지 않았다. 위의 고정빈도와 고정 시간에서 유의한 차이가 나타난 것과는 대조적인데, 이는 평가자들이 글과 번갈아가면서 평가 기준표를 보는 특성을 반영하는 것이다. 즉 평가자들은 전체 평가 과정 중 글을 보는 횟수만큼 평가 기준표도 보지만, 글에 비해 짧은 시간동안 평가 기준표를 보며 그에 따른 고정 시간도 짧았다.

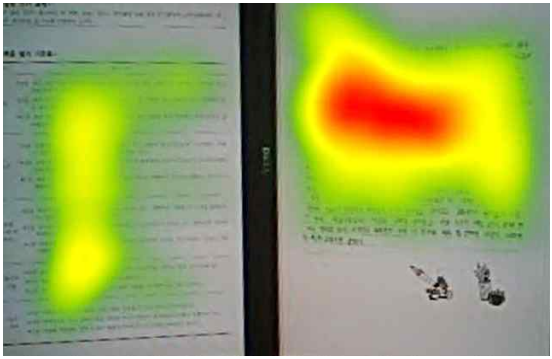

그림 2. AOI에 대한 heat map 분석(평가자 5)

<그림 2>는 눈동자 움직임의 고정 빈도를 기준으로 시각적인 분석을 제공하는 heat map으로 모니터 2(우)에 제시된 글에 대한 평가자의 고정 빈도의 차이를 관찰할 수 있다. 붉은색에 가까울수록 높은 고정 빈도를 나타내므로, 평가자들은 좌측에 위치한 평가 기준표보다는 글을 더 많이 보았으며, 특히 평가 기준표는 채점 중 모든 기술어를

읽기보다는 각각의 항목을 보는 데 활용됨을 확인할 수 있다.

2. 채점 수 누적에 따른 눈동자 움직임의 변화

일반적으로 글의 길이가 길어짐에 따라 평가자들은 글의 길이가 길거나, 글에 대한 판단을 내리기 어려울 때 평가 기준표를 더 많이 오래 보았다. 또한 평가 시간은 평가자에 따라 매우 다양한 범위로 나타난다. 이에 따라 보다 객관적인 수치를 관찰하고 결과를 일반화하기 위하여 채점 횟수 누적에 따른 눈동자 움직임 데이터의 분석에서는 한 편의 글을 볼 때의 평가 기준표를 보는 고정 빈도와 시간의 변화를 상댓값으로 처리하여 분석하였다.

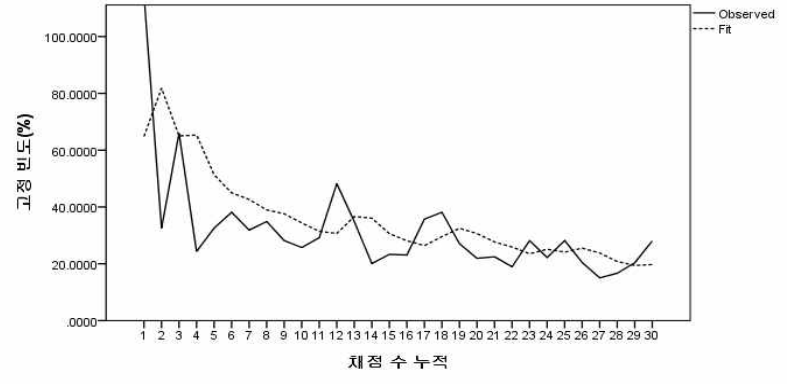

그림 3. 채점 수 누적에 따른 평가자 고정 빈도 평균의 추이

<그림 3>은 학생 글 채점 시기에 따라 글에 대한 평가 기준표의 고정 빈도의 상댓값을 그래프로 나타낸 것이다. 고정 빈도의 그래프는 '추세가 있는 시계열 그래프'의 형태로 나타나는데, 일반적인 꺾은선 그래프의 변화폭이 심하여 그래프의 추세를 알 수 있도록 시계열 데이터에서 추세 조정 지수 평활법(trend adjusted exponential smoothing

method)을 활용하여 관찰값의 불규칙 요인을 조정하여 적합값을 산출하였다. 그래프의 추세선(점선)은 글의 채점이 지속될수록 평가 기준표를 보는 평가자들의 고정 빈도에서 지속적인 하향 추세를 보여준다. 실선은 관찰값을 점선은 적합값을 나타낸다. 특히 평가자들이 10번째 글을 채점하는 시점까지 고정 빈도는 급격하게 감소하는 추세를 보이다가 이것이 점차 둔화되는 경향을 보이는 것을 관찰할 수 있다.

고정 시간의 상댓값 역시 <표 3>의 기술통계에서 보는 바와 같이 고정 빈도와 거의 차이가 없는 추세로 나타났다.[5] 즉 <그림 3>의 그래프 변동과 거의 유사하다. 이에 고정 시간은 전체 평가자의 평균이 아닌 각 평가자들의 개별 데이터를 보고하였다.

표 3. 채점 수 누적에 따른 평가 기준표 고정 빈도와 시간의 상댓값

채점 수	고정 빈도			고정 시간		
	N	M	SD	N	M	SD
1	10	115	105	10	114.7	104.7
2	10	32.5	36.5	10	32.5	36.51
3	10	65.9	97.6	10	65.91	97.52
4	10	24.3	26.8	10	24.33	26.85
5	10	32.6	33.6	10	32.56	33.55
6	10	38.1	37	10	38.12	37.05
7	10	31.9	38.1	10	31.83	38.1
8	10	34.9	49.4	10	34.87	49.39
9	10	28.2	22.5	10	28.17	22.5
10	10	25.7	23.6	10	25.73	23.58
11	10	29.2	32.1	10	29.17	32.03
12	10	48.2	59.8	10	48.24	59.82
13	10	34.9	34.5	10	34.86	34.44
14	10	20.1	19.3	10	20.05	19.29
15	10	23.3	27.6	10	23.29	27.6
16	10	23.1	24.3	10	23.04	24.28
17	10	35.7	37.2	10	35.73	37.23

5 고정 빈도와 고정 시간의 상댓값에 거의 차이가 없는 것은 눈동자 움직임 데이터가 동일한 시간 해상도(30Hz)로 측정되기 때문에 상댓값을 취할 경우, 시간에 따른 모든 빈도의 데이터가 수렴되기 때문이다.

18	8	38.1	33.2	8	33.27	33.88
19	8	27.1	23.2	8	32.44	26.42
20	8	21.9	20.7	8	21.95	20.71
21	8	22.5	19.4	8	22.51	19.44
22	8	19	18.9	8	18.96	18.88
23	8	28.2	28.7	8	28.12	28.65
24	8	22.2	21.7	8	22.18	21.7
25	8	28.2	30.9	8	28.16	30.82
26	8	20.5	23.2	8	20.46	23.26
27	8	15.1	14	8	15.04	13.96
28	8	16.7	19.7	8	16.65	19.64
29	8	20.4	24	8	20.32	23.95
30	8	28	35.4	8	27.95	35.24

<그림 4>는 개별 평가자들의 평가 기준표 고정 시간을 절댓값으로 나타낸 것이다. 이 연구에 참여한 평가자들은 한 편의 글을 평가할 때 빠르게는 15초 내외 길게는 3분 가까이 채점을 하였기 때문에 평가 시간의 범위가 커서 이에 따른 개별 평가자 간 절댓값의 편차가 크고 한 평가자 내의 변동도 심하긴 하지만, 대부분의 평가자들이 하향 추세를 보여주고 있다.

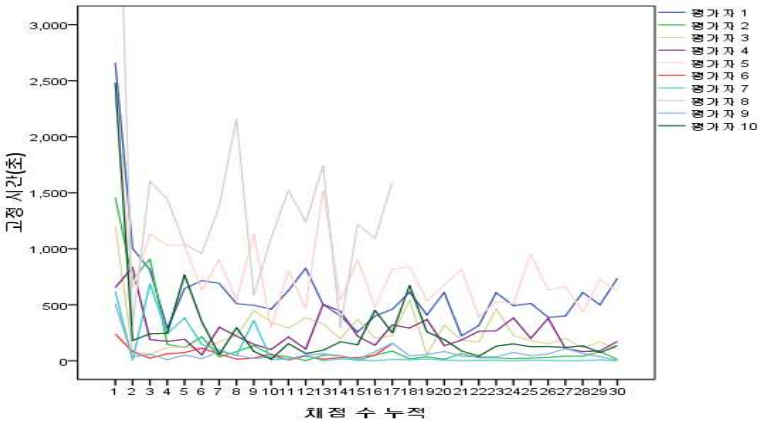

그림 4. 고정 시간 절댓값의 평가자 개별 데이터

또한 지속적인 하향 추세와 더불어 고정 빈도와 고정 시간에서 채점 초반부의 관찰값이 한 개를 걸러 들쭉날쭉한 경향을 공통적으로 보이는 것을 확인할 수 있다. 즉 주로 채점 초반부의 어느 시점에서 평가 기준표에 오래 시선을 머무른 이후에는 급격히 고정 시간이 감소하다가, 다시 상승하는 경향성이 관찰되고 있다. 이는 파지와 망각 효과에 의한 것으로 해석할 수 있다. 눈동자 움직임 데이터는 보는 과정에서 주의 집중이 일어나는 곳을 확인하는 데 유용한 도구이므로(Peterson & Beck, 2009: 582), 이것은 채점 과정에서 이루어지는 주의집중과 정보 처리 양상을 반영한다. 평가 기준표에 대한 표상을 아직 구성하지 못한 상태에서 이전의 글을 채점할 때 읽었던 평가 기준표에 대해 일부 구성된 표상에 의한 채점을 수행한 후, 이후의 채점에서 다시금 평가 기준표를 보는 눈동자의 움직임에 반영된다. 인간의 정보처리 과정에서 정보에 대한 정신적 표상을 기억에 유지하는 것을 파지(retention)라 하는데, 평가자들이 처음 읽은 평가 기준표는 일부만 표상으로 형성되고 나머지는 시간의 흐름과 새롭게 들어오는 글 정보들에 의해 망각된다. 파지 단계에 이르지 못한 정보들은 작업 기억 내에서 머무르게 되는데, 이것이 채점 과정에서 지속적인 반복을 거치면서 연습에 의한 파지가 지속적으로 행해지게 된다. 이후 평가자들이 고정하지 않고도 글에 대한 채점 수행이 일어난다는 것은 지속적인 평가 기준표의 적용에 따른 연습에 의한 효과로 평가 기준표에 대한 평가자 나름의 표상이 형성되었다고 해석할 수 있다. 따라서 채점 수가 누적되어 가면서 평가 기준표에 대한 고정 빈도와 시간은 감소한다. 이 연구의 평가자들이 보여주는 눈동자 움직임의 채점 초반부 경향은 고정 빈도의 감소가 파지와 망각에 따른 눈동자 움직임의 현상이 평가 기준표의 표상 형성을 설명하는 것을 뒷받침한다.

<그림 5>는 평가 기준표(AOI 1)에 대한 방문 빈도의 채점 수 누적에 따른 변화를 보여준다. 방문 빈도의 적합값의 그래프는 위에 제시된 고정 빈도와 고정 시간과는 달리 채점의 중반부인 10번째 글에서부터 급격하게 감소하는 경향이 지속되다가 채점 후반부에 완화되는 경

향을 보여준다. 이는 방문 빈도가 지속 시간과는 상관없이 AOI에 진입하였다가 벗어나는 순간을 1회로 계산하기 때문으로, 채점 초반부에 평가 기준표를 많이 보고 길게 머물렀던 경향이 고정 빈도와 고정 시간에서 나타났다면, 오히려 평가 기준표를 참조하여 평가하는 방문 빈도는 채점의 중반부에 비교적 더 많은 감소 추세를 보였다.

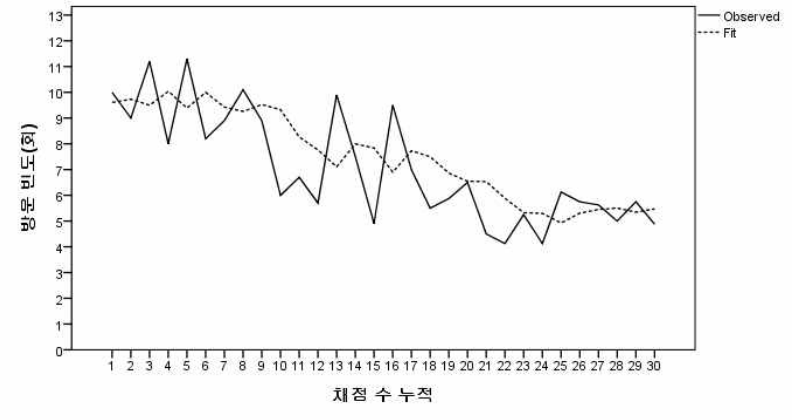

그림 5. 채점 수 누적에 따른 방문 빈도

3. 비선형 회귀 모형에 의한 예측값 분석

평가자들의 눈동자 움직임을 분석한 결과, 평가자들의 눈동자 움직임에는 채점 횟수의 누적에 따른 추세성이 있으며, 비선형적 형태의 특성을 보임에 따라 고정 빈도와 고정 시간에 대하여 비선형 회귀 모형을 활용한 예측값의 분석을 수행하였다. 이 연구에서 작성된 곡선 모형은 채점 수 누적에 따른 산포도에서 나타나는 고정 빈도의 급격한 수치의 감소로부터 일정치가 되는 기간을 거쳐 다시 완만한 값으로 바뀌어 마지막으로 어떤 값에 수렴하는 특징을 갖는다.

표 4. 채점 수 누적에 대한 비선형 회귀 분석 결과

Var	비표준화 계수		표준화 계수	t	p
	B	SD			
채점 수 누적	−15.876	2.862	−.724	30.760	.000
R^2(adj. R^2)=.523(.506), F=30.760					

여러 가지 비선형 회귀 모형의 결정계수를 비교한 결과 <표 4>와 같이 로그 함수 모형이 R^2이 0.5235으로 52.35%의 설명력을 가지며, 이에 따른 회귀 모형의 수식은 y = −15.88ln(x) + 71.21이다. 이 회귀 모형의 그래프는 <그림 6>과 같이 긴 꼬리의 형태를 띠는데, 이는 통계학적으로 인간 행동의 분포나 발생 확률에서 초반 기하급수적 감소 추세 이후 감소세를 유지하는 여러 현상을 설명하는 데 유용하다.[6] 이에 따라 발생 빈도의 대부분을 차지하는 좌측 소수의 부분과 발생 빈도는 낮으나 그 나머지 꼬리 부분에 해당하는 20%를 기점으로 하여 로그 함수 모형에 따른 예측값의 그래프의 채점 수를 산출하였다.

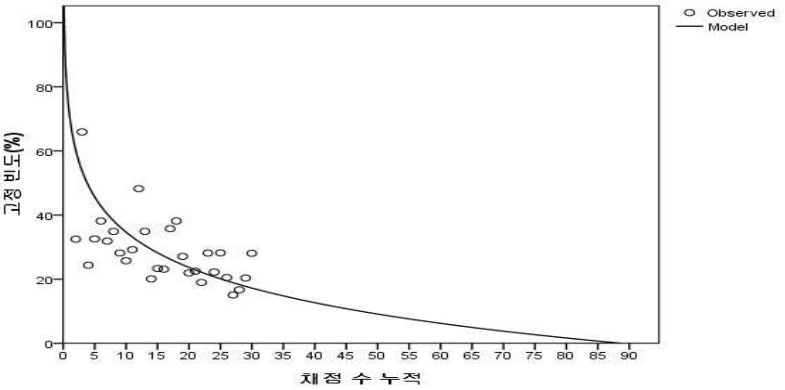

그림 6. 채점 수 누적에 따른 고정 빈도의 회귀 모형

6 멱 법칙(power law), 파레토 법칙(pareto −), 롱테일 법칙(The long tail −) 등이 이러한 유형의 그래프로 인간 행동을 설명하는 이론들이다.

로그 함수 모형에 따른 분석 결과, 고정 빈도의 상댓값이 20%가 되는 시점은 25번째 글을 채점할 때이고, 47번째 글을 채점할 때 고정 빈도가 10% 수준으로 감소하였으며, 88번째 글을 채점할 때 비로소 0이 되는 것으로 예측되었다. 고정 시간 역시 $y = -15.86\ln(x) + 71.17$, $R^2 = 0.5254$로 20% 지점은 25번 째 글을 채점할 때였으며, 10%가 되는 시점은 47번째 글이었고 88번째 글을 채점할 때 고정 시간의 상댓값이 0에 가까워지는 것으로 나타났다.

이에 따라 평가자들이 평가 기준표를 보는 행동의 소거 시점이 평가 기준에 대한 또렷한 표상이 형상된 시점으로 판단하여, 그 시점을 88번째 글을 채점할 때로 규정하였다. 이 결과는 쓰기 평가자 연수에서 사전 채점을 위한 적절한 학생 예시글 편수를 설정하는 데 도움이 될 수 있다. 결과적으로, 쓰기 평가 평가자 연수 기간에 사전 예비 채점을 위한 예시글을 88편 정도로 예측하여 프로그램을 설계할 필요가 있으며, 이를 위한 적절한 학생 글과 평가 기준표를 구성하여야 한다. 다만 평가 시행의 효율성, 목적과 상황을 고려하여 채점 수 누적에 따른 효과가 완만해지는 시점인 25편을 최소한의 채점 연습량으로 설정할 수 있을 것이다. 또한 평가자들은 채점 초반부에 평가 기준표를 읽고 해석하는 데에 많은 시간을 소요하는 것으로 나타났으므로, 평가자 협의 과정에서 평가자가 평가 기준표의 이해와 해석을 통해 명확한 내적 기준을 세우는 데 도움이 될 수 있는 방안을 마련하여야 할 것이다.

Ⅳ. 결론

쓰기 평가에 대한 눈동자 움직임 연구의 수는 매우 적지만, 눈동자 움직임이 제공하는 실시간의 데이터는 평가자의 복잡한 인지 과정을 밝힐 수 있는 많은 가능성을 제공하고 있다. 이 연구는 쓰기 평가자의 눈동자 움직임 데이터를 통해 평가 기준표를 중심으로 채점 수 누적에 따라 일어나는 변화를 관찰하였다.

연구 결과 첫째, 고정 빈도와 시간에서 통계적으로 유의한 차이가 나타났다. 평가자들은 전체 채점 과정에서 평가 기준표(AOI 1)와 글(AOI 2)을 보는 빈도와 시간에 유의한 차이가 있었으며, 방문 빈도에는 차이가 없었다. 이는 평가자들이 평가 기준표와 글을 번갈아 보는 특성을 반영하지만, 시간과 빈도에서 글의 우세를 보여준다.

둘째, 채점 수 누적에 따른 고정 빈도와 시간의 상댓값에서 평가자들은 지속적인 하향 추세를 나타냈으며, 특히 1번째 글에서 10번째 글을 채점할 때까지의 채점 초반부에 급격한 하향을 보였으나 이후 안정적인 하향 추세를 보여주었다. 또한 채점 초반부 관찰값의 불규칙한 변동은 평가 기준표에 대한 파지와 망각에 따른 지표로 해석되었다.

셋째, 채점 수 누적에 따른 고정 빈도와 시간의 상댓값을 통해 로그함수의 회귀 모형을 생성한 결과, 평가자들은 고정 빈도와 시간에서 88번째 글을 채점할 때에 평가 기준표를 보는 비율이 0에 가까워지는 것으로 나타났으며, 회귀 모형의 그래프에서 긴 꼬리를 형성하는, 즉 채점 수 누적에 따른 고정 빈도와 시간의 연습 효과가 미미해지는 시점은 25편을 채점할 때이므로 평가 시행의 효율성, 목적과 상황을 고려하여 최소 25편 이상을 채점 연습량으로 제공하여야 한다. 이는 평가자 연수와 평가자 협의 과정에서 샘플 채점 구성에 대한 논의점을 제공하였다.

이 연구의 데이터가 고등학교 2학년을 대상으로 한 설명문을 대상으로 하였다는 점, 평가 기준표는 평가 상황에 따라 달리 쓰일 수 있다는 점을 고려하면, 이 연구의 결과를 일반화하는 데에는 한계가 있을 것이다. 그러나 이러한 연구의 제한점에도 불구하고, 이 연구가 쓰기 평가의 평가자에 대한 인지 과정을 눈동자 움직임을 통해 정량적으로 측정하여 평가 과정에 대한 실증적 데이터를 제공하였다는 데 의의가 있다. 이 연구는 실제 평가 과정에 대한 실증적 연구로서, 실제 채점에 개입되는 다양한 변수에도 불구하고 아직 밝혀지지 않은 쓰기 평가 과정의 많은 부분을 탐색하기 위해 지속적으로 연구되어야 할 필요가 있을 것이다.

* 본 논문은 2015.11.02. 투고되었으며, 2015.11.03. 심사가 시작되어 2015.12.01. 심사가 종료되었음.

참고문헌

박영목(2008),『작문교육론』, 역락.

박영민(2015),「작문 평가의 평가자 신뢰도」,『국어학과 국어교육학』, 서울: 채륜.

박영민(2014),「손글씨 설명문과 워드 설명문 평가 과정에서 나타나는 국어교사 눈동자 움직임의 차이」,『국어교육학연구』49(2), 193-224.

박영민·김승희(2007),「쓰기 효능감 및 성별 차이가 중학생의 쓰기 수행에 미치는 효과」,『국어교육학연구』28, 국어교육학회, 327-359.

이정모 외(2002),『인지심리학』. 학지사.

Anderson, J. R. (1990). *Cognitive psychology and its implications*. WH Freeman/Times Books/Henry Holt & Co.

Andrade, H. G. (1997). Understanding rubrics. *Educational Leadership*, 54(4), 14-17.

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: Evaluating assessment essays. *College Composition and Communication*, 315-327.

Boyd, P., Ashworth, M., Bloxham, S., & Orr, S. (2009). Grading student work: Using think aloud to investigate the assessment practices of university lecturers. British Educational Research Association conference, Manchester, UK.

Colton, D. A., Gao, X., Harris, D. J., Kolen, M. J., Martinovich-Barhite, D., Wang, T., et al. (1997). Reliability issues with performance assessments: A collection of papers. *ACT Research Report Series,* 97-3.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86(1), 67-96.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, 5(1), 7-29.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. *Research on Writing: Principles and Methods*, 75-98.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a Multiple-Trait scoring instrument across ESL writing assessment contexts. *Language Learning*, 41(3), 337-373.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, 201-213.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective Vol. 3*. Peter Lang.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3(4), 331-345.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. *Performance Testing, Cognition and Assessment*, 92-114.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10), 71-81.

Paulson, E. J., & Freeman, A. E. (2003). *Insight from the eyes*. Heinemann Educational Books.

Peterson & Beck. Eye movement and memory. In Liversedge, S., Gilchrist, I., & Everling, S. (2011). *The Oxford handbook of eye movements*. Oxford University Press.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In Fairness and validation in language assessment: selected papers from the 19th language testing research colloquium, Orlando, Florida (129-152). Cambridge University Press.

Stahl, J. A., & Lunz, M. E. (1991). Answering the Call for a New

Psychometrics. *Rasch Measurement Transactions*, 5(1), 127.

Stahl, J. A., Lunz, M. E., & Wright, B. D. (1991, April). Equating examinations that include judges (multiple facets). In annual meeting of the National Council of Measurement in Education, Chicago.

Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15(2), 263-287.

White, E. M. (1984). Holisticism. *College Composition and Communication*, 35, 400-409.

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, 25, 37-53.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, 4(1), 83-106.

Wolfe, E. W. (2005). Identifying rater effects in performance ratings. *Performance appraisals: A critical view*, 91-103.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, 15(4), 465-492.

초록

채점 수 누적에 따른 쓰기 평가 과정의 눈동자 움직임 연구

이지원 · 박영민

이 연구는 눈동자 움직임 추적을 통해 쓰기 평가에서 평가 기준표와 글을 AOI로 설정하고 평가 기준표의 채점 수 누적에 따른 고정 빈도와 시간의 소거 과정을 탐색하였다. 첫째, 평가자들은 전체 채점 과정에서 고정 빈도 약 25%, 고정 시간 약 28%의 비율로 평가 기준표에 눈동자를 고정하였으며, 글에는 41.4초, 평가 기준표에는 16.2초의 고정 시간의 분포를 보였다. 둘째, 채점 수 누적에 따른 30편의 글 채점 과정에서 평가자들은 평가 기준표의 고정 빈도와 고정 시간의 지속적인 하향 추세를 보였으며, 특히 채점 초반부인 10번째 글까지의 급격한 하락 이후 완만하게 감소하는 경향을 보였다. 셋째, 채점 수 누적에 따른 고정 빈도와 고정 시간의 상댓값의 데이터를 바탕으로 회귀 모형을 작성한 결과, 로그 함수 모형이 가장 높은 설명력을 갖는 것으로 나타났으며, 예측값에 따른 평가 기준표의 고정 비율이 0이 되는 시점은 88번째 글인 것으로 추정했다. 다만 평가 시행의 효율성, 목적과 상황을 고려하여 채점 수 누적에 따른 연습 효과가 감소되는 시점인 고정 빈도의 20%에 해당하는 채점 수 25편을 최소한의 채점 연습량으로 설정할 수 있을 것이다.

핵심어 안구운동, 시선추적, 시계열, 작문 평가, 루브릭, 쓰기 채점, 평가자, 채점자, 예시문

ABSTRACT

Investigating Rater's Cognitive Processes in the Writing Assessment

—An Eye-Movement Study

LEE, Ji-Won & PARK, Young-Min

The purpose of this study was to investigate online processes of assessing writing through rater's the eye-movements. Rater's eyes-movements are measured during rating essays, and then their fixation counts and fixation duration between a rubric and essays are analysed that change as a time-series. Our data suggests that raters showed about 58% of fixation counts and 28% of fixation duration in rubric, and they fixed during whole rating processes 41.4 seconds at essay(AOI 2) and 16.2 seconds at rubric(AOI 1); fixation counts have been in sharp decline until they read the 10th essay, and then changed gentle decline trends following. We made a regression model of the log function base on relative fixation count and fixation duration data. We estimated a 88th essay as the predicted value that fixation rate is a zero. This means a time point that raters construct a cognitive representation of the rubric. And also we discussed that practices should require about 25 minimum rating sample essay for rater training that start long-tail trend line on our non-linear regression. This means a practice effect is slight that is represent the inclination declines on the graph.

KEYWORDS rubric, eye-tracking, time series, rater, anchor paper,

representation of criteria, fixation, rating, scoring, evaluating, judgement, decision-making, cognition, non-linear regression model