Modern English Education, Vol. 10, No. 3, Winter 2009

영작문 평가를 위한 채점자 훈련의 방향

이 지 연
총신대학교

Yi, Jyi-yeon. (2009). How should English teachers as raters be trained for writing assessment? *Modern English Education, 10*(3), 217-241.

This paper aimed to find what English teachers at secondary schools in Korea did during their rating assessment and to suggest what effective rater training session/programme for them should be like. Three English teachers at various secondary schools participated as raters for this study, and they were asked to rate six pieces of writing samples written by secondary school students, using FCE rating scale for writing assessment. They were also supposed to do think-aloud while rating so as to reveal their rating behaviour, and then their protocols recorded on cassette tapes were transcribed. The results showed that; the raters did not prevent themselves from relying on their subjective criteria; they failed to stick to the scale throughout their scoring; they were likely to rate a sample, comparing with preceeding samples; they often had difficulties in understanding some descriptors within the scale; the very first impression which a sample made on a rater often affected on his/her rating; they sometimes showed halo effect in their rating. From these findings it can be suggested that rater training session or programme for English teachers should be devised, reflecting their rating behaviour and help them improve test validity, and further intra- and inter-rater reliability.

[writing assessment/rating/rater training/쓰기 평가/채점/채점자 훈련]

I. 서론

영어가 세계 공용어로서 기능함에 따라 영어 사용 능력의 요구가 확대되어 왔고, 근래에는 이에 더하여 영어 듣기나 읽기와 같은 수용 기술에 대한 능력 뿐 아니라 말하기 및 쓰기 기술과 같은 발화 기술에 대한 요구 수준도 높아져 있는 것이 전 세계적인 추세이다. 교육에서 이와 같은 추세는 평가 영역에서도 그대로 적용되어 나타나고 있다. Test of English as a Foreign Language(TOEFL)와 같은 표준화 시험에서도 영어의 발화 기술에 대한 직접

평가(direct testing) 제도가 도입되고 있고, 한국에서도 영어 말하기 및 쓰기 기술에 대한 교육 및 이에 대한 평가가 초등학교부터 중등, 대학 수준에 이르기까지 여러 레벨에서 시도되거나 이미 이루어지고 있다. 초등학교에서는 영어가 교과과정에 처음 도입되는 시기이므로 의사소통 중심의 말하기 및 듣기 기술 발달에 초점을 두면서 공교육에서는 쓰기 기술이 고학년 이후에야 처음 도입되지만, 중등학교 수준에서는 말하기 및 쓰기 기술이 수행 평가의 일환으로도 다루어지고 있고 대학에서도 영작문 및 영어 말하기 기술은 어느 학교에서나 공통으로 이루어지는 교과들 중에 속한다. 우리나라의 교육 방향 및 흐름에 중대한 영향을 미치고 있는 대학 입학시험에서는 말하기나 쓰기 기술이 아직 포함되어 있지 않지만, 특별 목적 중·고등학교의 입학시험에는 지원자들의 영어 구사 능력을 판단하기 위한 중요 항목으로 영어 말하기 기술이나 쓰기 기술이 포함되어 있는 실정이다. 특히, 2009학년도부터는 영어과 중등교원선발시험이 3단계로 이루어지는데 그 중 2단계와 3단계에서 각각 영작문 능력과 영어 말하기 능력평가가 포함되는 실정이다.

이처럼 영어 발화 기술에 대한 교수 및 평가가 이전에 비해 활발히 이루어지게 되었는데(김형엽, 2006; 박상옥, 이유진, 2009; Sookyung Cho, 2009) 발화 기술 중 본 연구의 주제와 관련하여 볼 때 쓰기 기술에 대한 평가에 초점을 맞추어 논하도록 하겠다.

쓰기 기술에 대한 평가가 이루어지기 시작하면서 이에 대한 관심을 기울이게 되었는데, 쓰기 기술은 일반적으로 듣기나 읽기에 대한 평가보다 더 어렵다고 여겨진다. 이는 평가의 채점 단계에서 그 이유를 찾을 수 있다. 일반적으로 평가는 3단계로 구성된다고 본다. 평가의 개발, 시행 및 채점이 그것인데, 전자인 쓰기 기술과 후자인 읽기 및 듣기 기술의 평가를 구분지어 주면서 쓰기 기술 평가에 어려움을 야기하는 가장 큰 특징은 이 중 채점[1] 단계에 있다(Hamp-Lyons, 1990). 왜냐하면, 읽기나 듣기 기술의 평가에는 선다형 유형(multiple-choice test)의 문제가 가능하여 채점상의 실용도와 신뢰도를 높일 수 있는 반면에, 쓰기 기술에 대해서는 이와 같은 평가 유형이 적절치 않으며, 설령 그런 유형으로 평가할 경우 간접 평가가 됨으로써 평가의 내용 타당도가 떨어지게 된다. 이와 같은 선다형 유형을 통한 간접 평가보다는 직접 쓰게 하는 직접 평가로 시행하는 것이 평가의 구성 타당도(construct validity)를 고려할 때 바람직하다.

쓰기 평가에 있어서 이와 같은 채점 상의 절차로 인해 야기되는 어려움들이 있다. 첫째, 기계로 채점하지 못하고 사람이 일일이 채점하게 되기가 쉬우

1 '채점'이라는 의미는 실제로 점수를 매기는 과정을 말하고 '평가'라는 용어는 평가 문항의 개발과 실제 평가 시행, 채점 등을 포괄하는 좀 더 광의의 의미로서 사용되지만, 본 연구 내에서 경우에 따라서는 이 둘 사이에 의미의 차이 없이 혼용하여 쓰기로 한다.

므로[2] 다수의 채점자들을 고용하기 위해 예산이 더 필요하게 되고, 둘째, 채점하는 데에 있어서 각각의 채점자들이 평가하는 동안 채점자 내 신뢰도(intra-rater reliability)를 유지하지 못할 가능성도 있으며, 마지막으로, 각 채점자의 직업적 배경, 성격적 배경, 채점 경험에 있어서의 배경 등에 있어서 다양하므로 그로 인해 채점자들 간의 신뢰도(inter-rater reliability)를 유지하지 못할 수도 있게 된다. 채점자 간 신뢰도나 채점자 내 신뢰도의 문제는 단순히 채점에 있어서의 일치도/일관성 측면이라고 볼 수도 있지만 이러한 문제는 더 근본적으로 채점자가 채점에 있어서 구성 타당도를 가지고 있느냐의 문제에서 연유되는 것일 수 있다(Weigle, 1998). 즉, 무엇을 평가해야 하는 지, 평가하고자 하는 항목들이 각각 무엇을 뜻하는 지, 각 레벨의 요구 수준은 어느 정도인 지 등에 대해서 채점자가 분명하게 이해하고 있지 못하고 있기 때문에 파생될 수 있다는 것이다. 어떠한 이유에서 연유된 것이든지 간에 채점자 간 혹은 채점자 내 신뢰도를 유지하지 못하게 되면 그 결과, 어떠한 배경을 지닌 채점자에 의해 채점되느냐에 따라서, 또는 동일한 채점자에 의해서라도 어떤 상황에서 채점되느냐에 따라서 일관성이 떨어지는 채점 결과가 나올 수 있게 된다. 그런데 교실 내 평가가 아닌 표준화 시험은 평가의 결과가 피시험자의 진로에 큰 차이를 초래할 수 있게 되는 고부담 평가(high-stakes testing)이기가 쉬우므로 이러한 문제는 매우 신중히 고려해야할 문제가 된다.

이에 대한 대처 방안으로서 다양한 방법이 가능한데, 그 평가에 관여된 모든 채점자들이 채점할 때 단순히 주관적 평가 기준에 의해 평가하는 것이 아니라 공통된 평가 척도를 사용하게 한다거나, 채점자들을 모두 모이게 하고 채점자 훈련 세션을 갖거나, 각 채점자들이 개별적으로 채점자 훈련을 할 수 있는 훈련 키트를 사용하는 방법 등을 그 예로 들 수 있겠다. 평가 척도를 사용하게 되면 그 안에 평가 항목 및 각 등급의 수준이 명시되어 있으므로 이러한 평가 척도 없이 채점자의 주관적/인상적 판단에 따라 채점하는 경우에 비해서는 채점의 효율성을 높일 수 있게 되므로 근래에는 분석적 평가 척도(analytic rating scale)이든 통합적 평가 척도(holistic rating scale)이든지 간에 평가 척도를 사용하는 것이 더 장려되는 추세에 있다(Alderson, 1991). 한편

---

[2] 이를 위한 한 방법으로 기계 채점을 이용하는 것을 들 수 있을 것이다. Warschauer와 Ware(2006)는 1960년대 이래로 개발되어 사용되어 오고 있는 Project Essay Grade, the Writer's Workbench, My Access!, Criterion and the Intelligent Essay Assessor와 같은 다양한 기계 채점 체계들을 논의하고 있다. 이와 같은 기계 채점을 통해 즉각적이고도 개인별의 다양한 평가와 피드백을 제공할 수 있다는 장점이 있기도 하지만, 이와 같은 프로그램들의 유용성에 대해서는 아직 회의적이다. Warschauer와 Ware(2006)가 지적하고 있듯이, 이와 같은 연구들은 이런 프로그램들을 개발한 회사에 의해 실시되고 있을 뿐 어떤 연구 논문으로 발행된 경우가 없으며, 이와 같은 프로그램들에서는 철자나 글의 분량과 같은 눈에 띄기 쉬운 부분에 대한 채점이 전체 점수에 영향을 크게 미치는 것으로 보아 평가의 타당성에 대해서도 검증되지 못했다.

채점을 시작하기 전에 집단적 혹은 개별적 채점자 훈련 세션을 두어 채점자들이 훈련받게 되면 자신의 채점 상의 문제점이나 특징에 대한 인식을 높일 수 있게 되어 이것이 채점자 내 혹은 채점자 간의 신뢰도를 높일 수 있게 된다고 보고 있으며 이에 대한 연구들(A. Brown, 1995; Freedman, 1981; Kondo-Brown, 2002; Lumley, 1995; Weigle, 1994, 1998)도 많이 이루어져 있다.

그런데 채점자 훈련 단계를 가지려면 시간적, 비용적, 자원적인 면에서 현실적인 어려움이 예상된다. 따라서 채점자 훈련 단계를 실제로 가지려면 이를 효과적으로 운영하기 위한 방안을 강구하는 것이 필요하게 된다. 즉, 채점자 훈련을 실시하되 어떠한 측면에 초점을 맞추어 실시해야 그 효과를 높일 수 있을 지를 연구하여 시행하는 것이 필요하다. 그러기 위해서는 채점자들의 채점 과정에서 나타나는 양상들을 먼저 파악하여 이를 바탕으로 훈련의 목적 및 내용을 정하는 것이 효과적일 것이다. 그런데 국내 영어 교육 상황에서 채점자 훈련의 효율성을 높이기 위한 방안에 관한 연구들(이영식, 2000; 최연희, 2002)은 그리 많지 않은 실정이다. 따라서 본 연구는 쓰기 기술에 대한 평가에 있어서 채점자들의 채점 과정에 나타나는 양상들이 어떠한 지를 알아보고 효과적인 채점자 훈련의 방향을 제시하고자 한다.

II. 이론적 배경

1. 채점자 훈련의 필요성과 채점자들의 배경

채점자 훈련이 필요한 것은 일차적으로 채점자들 간의 차이가 존재하기 때문이다. 채점자들은 공통적으로 평가 척도를 사용하며 평가한다 하더라도 평가에 있어서 차이를 보일 가능성이 매우 높다. 그것은 각 채점자들은 평가하는 과정 중에 자신이 가지고 있는 고유한 배경을 가지고 오기 때문이다. 그 결과, Weigle(2002)이 지적하듯이 채점자들이 공통으로 사용하는 평가 척도에 대해서도 그에 대한 해석과 적용이 다를 수 있다.

평가에 영향을 미치는 채점자의 배경은 다양한 측면들에서 그 원인을 찾을 수 있다. 채점자의 직업적 배경이 무엇인 지, 채점자가 목표어의 모국어 화자인 지 아닌 지, 채점자의 성격적 특성은 어떠한 지 등이 대표적인 측면들이라고 할 수 있다.

첫째, 채점자의 직업적 배경이 채점자들의 평가에 영향을 미칠 수 있다. 이와 관련하여 이루어진 연구들의 주요 이슈는 언어 관련 직업을 가진 채점자 집단 대 언어와 관련이 없는 분야에 종사하는 채점자 집단 간에 평가 특성의 비교이다. 예를 들어 Lumley(1995)는 그의 연구에서 의료진들의 의사소통적

수행 평가를 채점해야 하는 상황에 있어서 의사 채점자 집단과 ESL 교사로서의 훈련을 받은 채점자 집단의 평가를 서로 비교하였다. 그 결과, 이러한 시험에서 ESL 채점자들이 의사 채점자들을 대신할 수 있으리만큼 두 집단이 평가 결과에 있어서 높은 일치도를 보였다.

이 밖에도 이 주제와 관련하여 많은 연구들(J. D. Brown, 1991; A. Brown, 1995; O'Loughlin, 1992)이 행해졌는데 그들의 연구 결과를 살펴보면 채점자 집단들이 그들의 직업적 배경과 상관없이 그들이 부여하는 최종 점수는 서로 비슷한데 각각의 평가 항목에 대한 개별 점수에 있어서는 집단 간에 차이가 있다는 것이 공통된 결론이다. 그리고 이러한 불일치는 평가하는 동안에 채점자의 관심에 있어서의 차이 및 사용하는 평가 척도에 대한 인식의 차이로 인해 발생할 수 있다고 설명하고 있다.

그 중 J. D. Brown(1991)은 영어과 교수들로 구성된 채점자 집단과 ESL 교수들로 구성된 채점자 집단들을 대상으로 연구하였다. 그 채점자들은 120개의 작문에 대해서 통합적 평가 방법으로 평가를 하였는데, 이들이 채점한 작문 샘플의 1/2은 ESL 학생이 쓴 것이고 나머지 반은 영어 모국어 화자 학생이 쓴 것이었다. 이 연구에서 채점자들은 평가하면서 응집, 내용, 철자법, 짜임새, 통사론 또는 어휘와 같은 측면에 대해서 각 작문의 우수한 측면과 그렇지 못한 측면을 규명하도록 지시 받았다. 이 두 피시험자 집단은 전체 평균 점수에 있어서는 유의미한 차이가 없었다. 더욱이 두 채점자 집단 모두 동일하게도 각 작문 샘플의 가장 우수한 측면으로서 '내용' 측면을 지적하곤 했다. 하지만 이 두 채점자 집단 간에는 차이가 있었는데, 영어과 교수 집단은 ESL 교수 집단에 비해서 '응집성'과 '통사적 구조' 측면에 대해 더 많은 주의를 기울였으며, ESL 교수 집단은 영어과 교수 집단에 비해 '짜임새'에 대해 더 많은 주의를 기울였다. 한편, 각 작문 샘플에서 가장 부족한 측면들을 규명해 내라는 요구에 대해서 두 채점자 집단 모두 '통사적 구조'를 가장 빈번히 선택하였는데 영어과 교수 집단은 '철자법'에 대해서 ESL 교수 집단보다 더 자주 주의를 기울였고, ESL 교수집단은 '내용'에 대해서 영어과 교수들보다 더 주의를 기울였다. 이러한 차이로 볼 때 채점자들이 평가할 때 동일한 점수를 준다 하더라도 그들의 직업적 배경에 따라서 다른 관점에 근거하여 그런 결과를 내는 것을 알 수 있다.

O'Loughlin(1992)은 영어를 모국어로 사용하는 학생들을 가르치는 교사 집단과 영어를 제 2언어로 사용하는 학생들을 가르치는 집단의 평가를 비교하였는데, 이 두 집단들은 분석적 방법과 통합적 방법을 통합하여 사용하는 경우에서보다 통합적 평가 방법만을 사용할 때에 평가 상 더 높은 일치도를 보이는 것으로 나타났다. 이것은 두 집단이 분석적 평가를 하는 상황에서 집단 간에 차이를 보인다는 것을 나타내 준다. 또한 이 집단들은 평가의 엄격성에 있어서도 차이를 보여서 영어를 모국어로 사용하는 학생들을 가르치는 교사

집단은 L2로서 영어를 배우는 학생들을 가르치는 교사 집단보다 분석적 방법과 통합 방법을 함께 적용하여 평가할 때 평가에 있어서 더 엄격했다. 이로부터 O'Loughlin은 결론짓기를, 두 채점자 집단은 평가 양상에 있어서 서로 차이를 보이며, 통합적 평가 방법은 "이 두 집단 간의 중요한 차이를 드러내지 못할 수 있다"(p. 39)고 결론짓고 있다.

A. Brown(1995) 역시 이 문제를 연구했는데 이 연구에서는 일본어를 목표어로 학습하는 경우를 다루었다. 일본어로 관광안내자로 구성된 채점자 집단과 외국어로서 일본어를 가르치는 일본어 모국어/준모국어 화자들로 구성된 채점자 집단을 대상으로 연구를 하였다. 이들은 비디오에 녹화된 51명의 말하기를 평가하였는데, 평가할 때 언어적 기술 측면과 과업 이행 정도를 평가하도록 하였다. 이들의 평가 결과는 전체적으로 서로 비슷한 것으로 나타났다. 하지만, 언어의 특정 자질을 인식하는데 있어서는 집단 간에 차이를 보였다. 일본어를 가르치는 교사 집단은 평가할 때 언어적 측면에 대해서 보다 엄격한 측면이 있었던 반면에, 일본어 관광 안내자 집단은 이에 대해 좀 더 관대하게 평가하는 경향이 있었다. 또한 평가 척도의 사용에 있어서도 차이를 보였는데, 교사 집단이 관광가이드 집단보다 최저 점수와 최고 점수를 부여하기를 더 꺼려했다.

둘째로, 채점자가 목표어의 원어민 화자인지의 측면이 평가에 영향을 미칠 수 있다. 이러한 측면에 대한 연구들은 평가 척도를 사용할 때의 차이점 및 엄격성에 있어서의 차이에 연구의 초점이 맞추어져 왔는데, 예를 들어, Hill (1997)은 인도네시아의 영어 능숙도 평가 내의 쓰기 부분 평가를 대상으로 이를 연구를 하였다. 이 연구에는 영어 모국어 화자(호주인) 집단과 비모국어 화자 집단(인도네시아인)이 각기 채점자 집단들로서 참여하였고, 이들은 각자 자신의 주관적 평가 기준을 사용하여 평가하도록 한 후 이들의 평가를 비교하였다. 이들의 평가에 있어서의 엄격성과 일관성을 조사해본 결과, 두 집단 간에 차이점이 발견되었다. 모국어 화자 채점자들은 일관성과 응집성 측면에 대해서 두 등급 간에 걸쳐 있는 양상을 보이는 글을 평가할 때 비모국어 화자 채점자들보다 더 엄격하게 평가하는 반면, 비모국어 화자 집단은 최상의 등급을 주는 데에 있어서 모국어 화자 채점자 집단보다도 더 인색한 것으로 나타났다. 이에 대해서 Hill은 비모국어 화자 채점자 집단은 최상의 등급에 대해 영어 모국어 화자의 능숙도 수준을 염두에 두고 이를 기준으로 평가하기 때문일 지도 모른다고 분석하였다. 또한, 영어 모국어 화자 채점자들이 비모국어 화자 채점자들보다도 평가 경력이 더 많았음에도 불구하고 비모국어 화자 채점자 집단이 모국어 화자 채점자 집단보다 채점자 간 신뢰도 뿐 만 아니라 채점자 내 신뢰도도 더 높은 것으로 나타났다.

셋째, 채점자의 성격이 어떠한 지도 평가에 영향을 미칠 수 있다. 이와 관련한 연구들은 주로 Myers-Briggs Type Indicator(MBTI)에 따라 채점자의 성격

유형을 분류하면서 연구하였다. 성격 유형과 평가 간의 상관관계에 관한 많은 연구들 중 Gowen(1984) 및 Jensen과 DiTiberio(1989) 등 이 분야의 연구들을 광범위하게 망라해놓은 Carrell(1995)의 연구가 주목할 만하다. Carrell(1995)의 연구에서는 여타 연구들에서처럼 채점자의 성격 유형과 평가 간의 관련성을 알아보고자 할 뿐만 아니라 작문 시험의 피시험자의 성격 유형과 평가 간의 관련성, 채점자의 성격 유형과 피시험자의 성격 유형과의 관련성까지도 분석하고자 하였다. 그 결과, 피시험자의 성격 유형에 있어서는 다른 축 보다도 외향-내성이라는 축이 평가에 가장 큰 영향을 미쳐서 내성적인 피시험자가 외향적인 성격의 피시험자보다 더 높은 점수를 얻는 경향이 있는 것으로 나타났고, 채점자의 성격 유형 또한 평가에 유의미한 영향을 미쳐서, 감각적이고 느낌의 성격을 지닌 채점자가 직관적이고 사고적인 채점자보다 더 높은 점수를 주는 경향이 있는 것으로 나타났다. 하지만, 채점자의 성격 유형과 피시험자의 성격 유형 간에는 관련성이 나타나지는 않았다.

이상에서 보듯이 채점자가 어떠한 배경적 특성을 가지고 있는 지는 평가에 영향을 미치게 된다. 이와 같은 사실은 평가 시 채점자의 엄격성이나 특정 평가 항목에 대한 편견(Kondo-Brown, 2002) 등을 이해하고 해석하는 데에 도움이 된다. 이처럼 채점자들은 자신의 다양한 특성들을 채점 과정에 가지고 오게 되고 그로 인해 채점자들 간에, 또는 각 채점자 내에 신뢰도를, 더 나아가서 타당도를 유지하기 어렵게 될 수 있게 된다. 따라서 앞서 서론에서도 언급한 것처럼 이러한 문제를 해결하기 위한 방안의 하나로서 채점자 훈련이 필요하게 된다. 그런데 실제로 채점자 훈련의 효과가 있는 지에 대해 다음 절에서 고찰해보기로 하겠다.

2. 채점자 훈련의 목표 및 효과

앞서 1절에서 살펴본 바와 같이 채점자들은 각기 다양한 배경을 가지고 평가에 임하게 되며 이러한 배경들로 인한 채점자들의 특징들은 그들의 평가에 영향을 미치게 된다. 이는 이들 평가의 채점자 간 또는 채점자 내 신뢰도에 영향을 미칠 수 있게 된다. 피시험자들의 수행이 비교적 공평하고 일관성 있게, 더 근본적으로는 타당도 있게 평가된다는 것을 보여주어야 평가의 권위와 객관성을 유지할 수 있다는 점을 감안할 때, 이 문제를 해결하기 위한 방안이 촉구된다고 할 수 있겠다. 그에 대한 해결 방안으로서 앞서 절에서 언급한 여러 방안들 중 채점자 훈련을 생각해볼 수 있다.

채점자 훈련은 여러 채점자들을 한 곳에 모이게 하고 훈련 세션을 가짐으로써 실행할 수도 있고, 이와 달리 각자 훈련을 할 수 있는 자료를 가지고 하는 자가 훈련을 통해서도 할 수 있겠다. 전자의 경우에서는 채점자들에게 평가 기준을 제시한 후 그것을 사용하여 몇 개의 샘플 작문들을 평가해보도

록 한 후 자신들의 평가 결과 및 채점 과정에 대해 다른 채점자들과 비교하면서 논의해 보도록 함으로써 평가 기준에 대한 이해 및 적용 방법을 공통적으로 하게 하려는 방식으로 진행된다. 한편, 자가 훈련에서는 채점자 훈련 키트를 가지고 실시할 수 있는데, 이 안에 들어있는 평가 척도와 평가 샘플을 가지고 각자 평가해본 후 그 평가 샘플에 대한 채점자들의 채점 사례, 평가 개발자의 평가 의도로 볼 때 바람직한 채점 방향에 대한 설명이 제시되어 있어서 이를 바탕으로 자신의 평가를 재조명해보는 식으로 진행될 수 있다.

이러한 채점자 훈련은 주로 채점자들 간에 채점 결과의 일치, 즉 채점자 간 신뢰도 향상을 위한 것으로 보일 수 있는데, 사실 평가에 있어서의 일치도/신뢰도 향상의 문제는 더 근본적으로는 평가의 타당도를 달성하기 위한 것이며 또한 그래야만 하는 것이라고 할 수 있다(Davies & Elder, 2005). 왜냐하면 채점자 간, 그리고 채점자 내에 있어서 채점의 일관성/신뢰도는 평가의 구성(construct), 평가의 목적 등에 대한 명확한 이해가 없이는 이루어질 수 없기 때문이다. 다시 말해서, 채점자 내, 그리고 채점자 간에 평가의 타당도가 달성되지 못하면 그 결과 일관성/신뢰도 있는 평가를 기대하기 어렵게 되고 그로 인해 평가에서 채점자 간의, 그리고 채점자 내의 신뢰도를 이루어낼 수 어렵게 되는 것이다.

이러한 채점자 훈련의 효과에 관한 연구들을 살펴보면 일부에서는 긍정적인 효과가 있다는 결론을 보여주고 있다. Freedman(1981)의 연구에서 보면 훈련을 통해 채점자들의 평가 행위에 변화가 나타났음을 볼 수 있다. Weigle(1994, 1998) 또한 자신의 두 번에 걸친 질적 또는 양적 실험 연구를 통해 채점자 훈련을 하게 되면 채점자들이 평가 척도를 그 척도에서 본래 의도된 대로 이해하고 채점하는 작문 샘플들에 대한 기대 수준을 조정하고, 동료 채점자들을 의식하게 됨(awareness)으로써 평가에 긍정적인 영향을 가져다주게 됨을 보여주고 있다. 하지만, 이것은 채점자 내의 신뢰도를 유지할 수 있게 해준다는 것을 뜻하지, 평가의 엄격성 면이나 다른 면에 있어서 채점자들 간의 차이가 완전히 사라지게 된다는 것을 뜻하는 것은 아니라고 하면서 채점자들 간의 차이가 완전히 사라지게 하는 것은 현실적으로 어려운 목표이며 채점자 훈련의 실제 목표가 될 수는 없다고 결론짓고 있다.

한편, 채점자 훈련의 효과가 의도한 수준에 이르지 못한다는 연구들도 있다. A. Brown(1995)은 일본어로 하는 관광 안내자 시험을 평가는 채점자들을 대상으로 연구하였는데, 이 연구에서 채점자가 일본어 모국어 화자이든 준모국어 화자이든지 간에 피시험자들이 수행 도중 보여주는 실수들에 대해서 관용해주는 수준, 즉 평가의 엄격성에 있어서 이 두 채점자 집단 간에는 여전히 큰 차이가 존재하는 것으로 나타났다.

이와 같은 연구 결과는 Kondo-Brown(2002)의 연구에서도 나타나고 있다. 즉, 연구 결과 채점자 훈련을 통해 채점자들의 채점자 내 일관성은 더 높아

지기는 했지만 채점자 간에 있어서는 여전히 큰 차이가 존재하고 있었다. 이러한 연구 결과들에 비추어 볼 때 채점자 각자의 평가에 있어서의 엄격성은 채점자 훈련 이후에도 여전히 지속되는 현상인 것을 알 수 있다. 이러한 현상은 Lumley(1995)의 연구에서도 나타났다. Kondo-Brown(2002)와 Lumley (1995)의 이와 같은 연구 결과들에 따르면 “채점자 훈련은 평가하는 데에 있어서 채점자 개개인의 전반적인 엄격성을 줄여주거나 완전히 제거해주지는 못한다… [대신] 이것은 채점자가 보다 자기 자신의 일관성을 유지해주도록 하는 데에 있어서 효과적이다”(Kondo-Brown, 2002, p. 4)라고 결론지을 수 있겠다.

이상을 요약하면, 채점자 훈련은 채점자 간의 차이를 제거하기 위한 것에 초점을 맞추는 것보다는 채점자 내의 일관성을 유지하게 해주는 데에 초점을 맞추는 것이 보다 현실적이라고 할 수 있다. 하지만, Weigle(1998)이 지적하듯이, 채점자 내 신뢰도에만 너무 많은 강조를 하는 것도 바람직하지는 못하다. 왜냐하면 채점자 내 일관성만 유지하려고 하다가는 구성 타당도에서 벗어나게 될 위험도 있기 때문이다. 이럴 경우 각 채점자 자신이 평가의 일관성은 유지할 수 있을지 몰라도 그 평가나 평가 척도를 통해 평가되고자 하는 것과는 다른 측면을 평가하고 있을 수도 있게 된다. 따라서 평가의 구성 타당도를 벗어나지 않는 범위 내에서 채점자 내 신뢰도를 유지하도록 노력해야 할 것이다.

III. 실험 연구[3]

1. 연구 대상

3명의 현직 영어 교사들이 채점자로서 본 연구에 참여했다. 표 1에서 보듯이 이들은 각기 다른 고등학교에서 영어 교사로서 근무하고 있었으며 그들 중 2명은 일반 고등학교에, 나머지 1명은 외국어 고등학교에 근무하고 있었다. 교사 근무 경력을 보면 각각 12년, 5년, 34년간의 교사 경력을 가지고 있었다. 그들 중 한 명은 여교사였고 나머지 2명은 남교사들이었으며, 그들의 연령을 보면 각각 40대 후반, 30대 초반, 60대 중반이었다. 이들은 모두 영어 관련학과 전공자였고, 대학원에서 영어교육 또는 영문학을 전공했다. 이처럼 본 연구에 채점자로서 참여한 교사들의 수가 적기는 하나 이들은 서로 다른

3 본 연구에서는 채점 과정 상 채점자들의 평가 행위에 초점을 맞추어 연구하고자 하였는데, 이들 채점자들이 사용한 평가 척도의 특성에 초점을 맞춘 연구를 보려면 본 연구와 동일한 연구 자료를 사용하여 수행된 이지연(2007)을 참조 바람.

배경을 가지고 있는 교사들이었다.

표 1
채점자의 배경

	채점자 1	채점자 2	채점자 3
성별	남	여	남
연령	40대 후반	30대 초반	60대 중반
교사경력	12년	5년	34년
근무학교	외국어고교 (경기 소재)	일반고교(서울 소재)	일반고교 (서울 소재)
학력/전공	대학원 박사/영어교육	대학원석사 /영어교육	대학원/영문

2. 평가 자료

채점자들이 본 연구를 위하여 채점할 평가 자료로는 경기도 소재 K 외국어 고등학교의 학생들로부터 수집한 영작문 샘플 중에서 6개의 영작문을 선택하여 사용하였다. 이 중 3개는 외국인 친구가 있다고 가정하고 그 친구에게 한국 내의 방문할 만한 곳에 대하여 소개하는 편지글의 과업이었고, 나머지 3개는 인터넷 사용의 장단점에 관하여 형식적 에세이(formal essay)를 쓰는 것이었다[4]. 이 샘플들은 박사 학위 논문(Jyi-yeon Yi, 2006)을 위하여 수집된 것인데, 수집 후 이 샘플들을 세 묶음으로 나누어 연구에 참여한 3명의 영어 교사들에게 한 묶음씩 나누어 주고 채점하도록 하였다. 이 때 작문 샘플들에 대한 교사의 평가를 알아보기 위하여 교사들에게 어떠한 척도도 사용하지 않고 교사 자신의 주관적 기준에 따라서 평가하도록 하였다. 이 때 최하위 등급을 1등급으로 하고 최상위 등급을 6등급으로 하여 6개의 등급 상에서 평가한다는 것 이외에는 어떠한 지시사항이나 안내도 주지 않았다. 교사들의 이와 같은 주관적 채점 결과에 따라 작문 샘플들을 각 등급별로 분류한 후, 편지글 과업의 글에서 1, 2, 3등급의 글을 한 편씩 추출하고, 에세이 과업의 글에서 4, 5, 6등급의 글을 한 편 추출하여 전체 6개의 작문 샘플을 수집하여 본 연구의 평가 자료로서 사용하게 되었다. 본 연구의 자료로서 6개의 샘플만을 사용하게 된 것은 채점자로 참여하는 교사들과의 협의 하에 정해진 것

4 작문의 유형을 비형식적인 글과 형식적인 글 두 유형으로 구분하고 각 유형의 대표적인 예로서 사적인 편지쓰기와 형식적인 에세이를 들었다. 그리고 모든 학생들이 이 두 가지 유형에 대하여 글을 쓰도록 함으로써 수집하는 작문 샘플에 있어서 글의 유형에 있어서 다양성을 확보하고자 노력하였다.

인데, 사용한 샘플의 수가 적다보니 본 연구에서 나오게 되는 연구 결과를 일반화시키는 데에는 한계점이 있을 수 있겠다.

3. 연구 방법

본 연구를 위하여 채점자로서 참여한 교사들은 First Certificate in English(FCE)의 영작문 평가를 위한 평가 척도를 사용하여 평가하면서 그 때의 채점 과정 동안에 머릿속에 드는 '생각을 소리 내어 말하기 과업'(think-aloud)을 하였다.

생각을 소리 내어 말하기 과업이란 질적 연구를 위한 방법들 중의 하나로서 특정의 과업을 수행하게 하면서 과업 수행 동안에 머릿속에 드는 생각을 모두 발화하도록 하여 과업 수행 과정 동안의 내적인 의식(consciousness)의 흐름을 관찰하고자 하는 방법이다(박경자 외 6인, 2001; 이지연, 2007). 어떤 과업을 수행하는 동안 자신의 머릿속에 드는 생각을 모두 발화한다는 것은 일반적인 상황에서는 수행되는 경우가 적으므로 매우 부자연스럽고 인위적인 과정이며 머릿속에 드는 생각들을 의도한 대로 실제로 모두 발화하는 지의 여부에 대해서 완전하지 못할 수 있으므로 이 연구 방법에 대한 비판(Glendinning & Howard, 2001)이 있기는 하나, 설문지나 면담처럼 과업이 수행된 후에 회고하는 것을 통해 기록을 얻는 것보다는 어느 정도는 충분하고도 즉각적이며 실제적인 기록을 얻을 수 있다는 점(Green, 1998)에 주목하면서 이 방법을 택하기로 하였다.

채점자로 참여한 교사들에게 평가할 때 사용하도록 제시한 평가 척도는 FCE에서의 쓰기 평가를 위한 일반적인 형태의 평가 척도였는데 FCE란 University of Cambridge Local Examination Syndicate(UCLES)에 의해서 1939년 처음 개발된 시험으로서 L2로서의 영어 능숙도 시험으로서 이를 통해 영어 읽기, 쓰기, 언어 사용, 듣기, 말하기의 5개 영역에 대하여 평가한다. 그 중 쓰기 기술은 2개의 작문을 하도록 요구하게 되는데 이들에 대한 평가 시 공통적으로 적용하기 위한 평가 척도 또한 개발되어 있다. 본 연구에서는 채점자들이 이 평가 척도를 사용하여 평가하도록 하였는데, 이 평가 척도는 중급 수준의 L2 영어 사용자들을 대상으로 하여 고안된 시험이므로 평가 자료의 수준들에 비교적 적절할 가능성이 높다고 여겨져 이 평가 척도를 선택하게 되었다.

3. 연구 절차

채점자들이 FCE 쓰기 평가 척도를 사용하여 6개의 영작문 평가 자료를 평가하기 전에, 사용하게 될 평가 척도 및 '생각을 소리 내어 말하기 과업' 이

두 가지 사항에 대해 사전 훈련 단계를 가졌다. 채점자들은 각각 연구자와의 1:1 면담을 통하여 연구자가 준비한 매뉴얼을 가지고 훈련을 받았다. 먼저, 이들은 FCE 평가 또는 FCE 평가 척도를 사용해본 적이 없으므로 FCE 평가에 대한 전반적인 소개를 먼저 다루고, 그에 이어 FCE 쓰기 평가 척도 내의 평가 항목 및 각 등급의 수준, 평가 척도 내의 용어들에 대해서 연구자로부터 설명을 들었다. 또한 '생각을 소리 내어 말하기 과업'에 대한 훈련 과정을 가졌다. 이 과업에 대해서도 채점자들이 실제 경험이 없으므로 이들은 이 과업의 목적 및 방법에 대해 소개한 매뉴얼을 가지고 연구자로부터 먼저 구두로 설명을 받았다.

평가 척도 및 평가 과업에 대해서 전반적인 소개를 들은 후, 본 연구를 위해 평가하게 될 6개의 평가 자료와는 별도로 연구자가 준비한 영작문 평가 자료 1개를 FCE 쓰기 평가 척도를 가지고 연습 과정을 밟았다. 본 연구는 이들 채점자들이 평가 척도를 사용할 때에 어떠한 평가 행위를 보이는 지를 관찰하여 이들에게 채점자 훈련을 제공할 때 효과적으로 하기 위해서는 어떠한 측면에 맞추어 해야 하는 지를 알아보기 위한 것이므로 본 연구에서의 채점자 훈련 과정은 FCE 평가 척도에 대한 이해를 돕기 위한 설명을 해주는 측면에만 초점을 맞추었지 그 외에 채점 과정 동안의 평가 행위에 대해서는 사전 안내를 하지 않았다. 또한 '생각을 소리 내어 말하기 과업'을 자연스럽게 수행할 수 있도록 하기 위해 이에 대한 훈련을 하는 것에 주안점을 두었다. 본 연구에서 얻고자 한 채점자의 평가 행위에 대해 자세하면서 정확한 자료를 얻기 위해서는 이들이 이 과업을 가능한 한 자연스럽게 수행해야 했다. 과업을 수행하는 동안 자신의 생각을 말로 드러내는 것을 어색하게 여기고 충분히 드러내지 못하는 경우가 발생하지 않도록 하기 위해 이에 대한 훈련에 보다 더 큰 초점을 맞추었다. 이들은 위에서 언급한 시험 평가 자료를 평가하면서 이들의 채점 과정을 말로 소리 내어 발화하도록 지시받았고 이들의 발화를 테이프에 녹음하였다. 평가와 녹음이 끝난 후 녹음된 내용을 연구자와 함께 들으면서 소감 및 개선해야 할 사항에 대해 나누었다. 이 과업을 수행할 때에는 채점자의 생각을 거침없이 모두 쏟아내도록 하는 것이 중요하므로 과업 수행은 우리말로 하도록 하였다.

이와 같은 채점자 훈련을 각 채점자와 개별적으로 가진 후 채점자들은 주어진 6개의 평가 자료를 FCE 쓰기 평가 척도를 사용하면서 평가하고 그 채점 과정을 모두 테이프에 녹음하였다. 본 과업은 연구자의 참석 없이 각 채점자 개별적으로 단독 수행하였으며 각 채점자로부터의 6개의 발화 자료를 수집한 결과 총 18개의 발화 자료를 수집하였다.

4. 분석 방법

채점 과업 수행으로부터 채점 결과 및 이들의 채점 과정에 대한 발화 자료가 수집되었다. 그 중 발화 자료가 본 연구의 목적과 관련되므로 이에 대해서만 분석하였다. 수집된 발화 자료를 연구자가 모두 전사(transcription)하고 나서 이들을 반복적으로 읽은 결과 이들의 채점 과정은 크게 두 가지 측면의 현상들로 나뉘는 것을 알 수 있었다. 그 하나는 평가 행위에 관한 측면들이었고, 나머지 하나는 평가 척도를 사용할 때의 평가 항목/양상(이지연, 2007)에 관한 것이었다. 이 중 본 연구는 평가 행위에 관한 측면들에 초점을 맞추어 이를 분석하였다. 그러기 위하여 나타난 현상들을 범주의 형태로 형성하고 이들 간의 위계도 고려하여 범주 체계[5]로 구성하였다. 이러한 범주 체계를 사용하여 발화 자료에 대한 양적 분석을 시도한 결과[6] 표 2와 같은 결과를 얻었다.

표 2
발화 자료 코딩에 대한 빈도 분석 결과

대범주	중범주	소범주	채점자 A	채점자 B	채점자 C	합계 (비율)
1. 채점 행위	1.1 글에 대한 첫 반응 보이기	1.1.1 분량에 관하여	4	3	3	10 (1.8%)
		1.1.2 분량 이외에 글 전체에 대하여	1	0	1	2 (0.4%)
	1.2 글 읽기	1.2.1 불분명하거나 어색한 어휘 또는 구를 다시 읽기	5	4	10	19 (3.5%)
		1.2.2 특정 목적을 염두에 두면서 글의 전체나 일부를 훑기	3	1	7	11 (1.8%)
		1.2.3 작문 과업에서의 문제를 언급하기	3	2	0	5 (0.9%)
	1.3 점수 부여하기	1.3.1 자신의 평가 기준에 따라 점수 부여 해보기	4	3	4	11 (1.8%)
		1.3.2 가능한 두 점수 사이에서 고민하기	27	18	13	58 (10.6%)

5 채점자들의 채점 행위 자체와 채점하는 측면들을 서로 나누어 분석 범주를 세움으로써 분석 범주의 전체적인 수를 줄이면서도 채점 과정동안 나타나는 모든 양상에 대한 분석이 가능하도록 하였다.

6 발화 자료 코딩 작업에 들어가기 전, 연구자의 코딩의 객관성에 대한 검증을 위하여, 각 채점자들의 발화 자료로부터 각각 한 개씩의 자료를 임의로 선택하여 이들에 대해서 연구자와 응용언어학 석사 과정생이 각자 코딩하였다. 코딩 결과를 비교하여 일치도를 확인해본 결과 일치도는 84.2%였다. 이 정도의 일치도를 바탕으로 하여 연구 결과를 해석하는 데에 크게 어려움은 없지만, 앞으로는 응용언어학에 관하여 좀 더 전문적 식견을 가진 연구자와 코딩 일치도를 검증해보는 것도 바람직하다고 여겨진다. 그럼으로써 코딩 일치도 및 연구의 신뢰도를 더 높일 필요가 있을 것으로 보인다.

	1.3.3 그 글에 대한 적절한 점수가 무엇일지를 고려하기 시작하기	30	29	32	91 (16.6%)
	1.3.4 점수 결정을 하기가 어려움을 나타내기	14	17	13	44 (8.0%)
	1.3.5 척도에서 기인된 점수와 자신의 주관적 기준에 의한 점수 사이에서 고민하기	9	6	8	23 (4.2%)
	1.3.6 주관적 기준에 의존하기	8	6	5	19 (3.5%)
	1.3.7 척도 언급하기	8	5	2	15 (2.7%)
	1.3.8 척도 내의 관련 부분을 그대로 소리 내어 읽기	6	7	9	22 (4.0%)
	1.3.9 척도에 대한 이해의 어려움 드러내기	17	19	7	43 (7.8%)
	1.3.10 척도에 대한 자신의 의견 나타내기	4	3	1	8 (1.5%)
	1.3.11 생각하고 있는 점수를 다시 체크하기	6	9	9	24 (4.4%)
	1.3.12 점수를 결정하기	34	35	34	103 (18.8%)
	1.3.13 주관적 기준에 의한 점수와 척도에 의한 점수를 비교하기	8	4	13	25 (4.4%)
	1.3.14 다른 글들에 대해 부여한 점수와 비교하기	4	5	3	12 (2.2%)
	1.4 채점자 개인적 견해 (채점자의 태도, 채점자의 유추)	1	1	2	4 (0.7%)
	소 계	196 (35.5%)	180 (32.6%)	176 (31.9%)	552 (100%)
2. 평가 항목 들에 관련 된 언급	2.1 짜임새	6	7	6	19 (5.1%)
	2.2 문단 구분	6	4	0	10 (2.7%)
	2.3 내용	6	8	6	20 (5.4%)
	2.4 어휘, 구: 세련, 적절성 여부	6	8	4	18 (4.8%)
	2.5 표현, 스타일	3	5	2	10 (2.7%)
	2.6 흐름/일관성	1	1	0	2 (0.5%)
	2.7 응집성	6	7	6	19 (5.1%)
	2.8 길이	4	4	3	11 (3.0%)

2.9 전체적인 언어 수준	5	6	6	17 (4.6%)
2.10 전체적인 문법적 정확성	6	4	6	16 (4.3%)
2.11어휘: 형태와 용법의 정확성	8	7	21	36 (9.7%)
2.12 철자, 구두점, 대문자쓰기	3	2	11	16 (4.3%)
2.13 그 외 개별적 문법 항목에 있어서의 오류	4	2	13	19 (5.1%)
2.14 문장 구조	17	14	34	65 (17.5%)
2.15 어색함	8	5	21	34 (9.1%)
2.16 의사 전달 가능성	3	6	4	13 (3.5%)
2.17 독자에 대한 관심 (target reader)	6	6	6	18 (4.8%)
2.18 언어 범위 (range)	8	7	6	21 (5.6%)
2.19 성실도	3	3	2	8 (2.2%)
소 계	109	106	157	372 (100%)

위의 표 2에서 보듯이 자료 분석은 '채점 행위'(대범주 1)에 관한 것과 '평가 항목들에 관련된 언급 및 채점'(대범주 2) 모두에 대한 분석 결과를 담고 있는데 이 중 대범주 2 '평가항목들에 관련된 언급'에 관한 항목들에 대한 논의(이지연, 2007)는 본 연구의 목적에서 벗어나므로 여기서는 논외로 하고 본 연구에서는 대범주 1 '채점 행위' 자체와 관련된 항목들에 관해서만 기술하도록 하겠다.

IV. 연구 결과 및 논의

발화 자료를 바탕으로 하여 귀납적으로 수립된 분석 체계에서 보듯이, 채점자들은 채점하는 동안에 글에 대한 첫 반응을 보이는 행위, 글을 읽는 행위, 점수를 부여하는 행위가 주로 관찰된 행위였다. 이러한 분석 체계를 바탕으로 한 코딩 결과를 논할 때 그 안에서 두드러지게 나타난 양상들을 중심으로 논하도록 하겠다.

코딩 결과를 바탕으로 보면, 채점 행위 중에서도 채점의 목적이나 문제 등을 염두에 두면서 글을 읽는 행위와 점수를 부여하는 행위가 더욱 두드러진

측면이라고 할 수 있겠다. 점수 부여하는 행위와 관련해서는 범주 체계에서 볼 수 있듯이 다음의 7가지 행위들이 나타났다. 즉, 1) 글의 첫 인상에 기반하여 머릿속에 처음 떠오르는 점수를 생각해보는 행위, 2) 평가하는 글에 대한 적절한 점수가 무엇일 지를 진지하게 고려하기 시작하는 행위, 3) 점수를 결정하기가 어려움을 표현하는 행위, 4) 척도 내용의 이해 및 적용에 어려움을 나타내는 행위, 5) 평가하는 글에 대하여 부여 가능한 두 개 정도의 점수들을 놓고 고민하는 행위, 6) 최종 점수를 내리기 전에 머릿속에 생각하고 있는 점수를 다시 한 번 체크해보는 행위, 7) 최종 점수를 결정하는 행위 등이 그것이다. 이 중 앞으로 채점자 훈련을 실시할 때 어떠한 점에 유의하여 해야 하는 지 본 연구의 목적과 관련하여 주목할 만한 특징 몇 가지들에 초점을 맞추어 결과를 논해보도록 하겠다.

첫째, 채점자들은 평가 자료를 평가하기 전에 먼저 그 글에 대한 첫 인상을 중심으로 반응하는 양상을 보이곤 하였다. 그 첫 인상은 아래 발췌에서 보듯이, 주로 글의 분량에 관한 경우가 많았다.

[채점자 A-01]
(생략) 음.. 이 학생은 우선 글이 너무 짧으네.. 뭐.. 짧긴 해도 내용은 비교적 정확하게 쓴 것 같은데… 뭐.. 내용이 충분하다고 볼 수는 없는 것 같고.. 뭐.. (생략)

[채점자 B-03]
(생략) 음... 우선... 너무 짧아서.... 너무 짧다보니까... 가볼만한 곳에 대해서 소개를 자세히 하지를 못했네... 어휘도... 뭐.... 제대로, 적절하게 사용된 것 같지 않고... (생략)

글의 분량에 대한 인상을 중심으로 하여 반응을 보이는데, 분량은 작문의 유창성을 평가하기 위한 하나의 측면으로 고려될 수는 있다(이지연, 2007). 하지만 사용하고 있는 평가 척도인 FCE 쓰기 평가 척도 내에는 이러한 측면들에 대한 언급이 없는 데에도 불구하고 채점자들이 이에 대해 주의를 기울이고 있다는 것에 주목할 필요가 있다. 왜냐하면 이것은 평가의 타당도에 영향을 미치게 되기 때문이다. 평가하는 글이 주는 첫 인상에 채점자가 주목하게 되면 그러한 첫 인상으로 인해, 그 평가 척도를 통해 평가해야 할 측면들에 온전히 주의를 다 기울이지 못하게 된다. 그렇게 된다면 표면적으로는 평가 척도를 사용하고는 있지만 실제로는 그 평가 척도를 통해 평가될 것으로 기대하는 방식으로 평가가 이루어지지 않으므로 평가의 타당도에 영향을 미칠 수 있게 되므로 채점자 훈련 시 고려해야 할 사항이 될 것으로 보인다.

둘째, 점수 부여 행위와 관련하여, 아래의 발췌에서 보듯이 사용하는 평가

척도에 대한 이해의 어려움을 보여준 행위들이 눈에 띄었다.

[채점자 B-01]
그 다음에 register와 format에 있어서는…항상 나는 이 항목이 뭐에 대한 건지 정확한 개념이 안 떠오르기는 하는데…뭐 특별히 편지나 친구한테 맞춰서 한 것은 아니고 그냥 무난하게 썼는데…이거에 대해서 합리적이라고 봐서 4라고 봐야 하는 건지, 3이라고 봐야 하는 건지…. 도대체 일관성이 없다는 것과, 합리적이라는 것과 3과 4의 차이가 뭔지… 그리고 5는 또 뭐라고 봐야 하는 건지.. 나는 이 항목에 대해서 늘 판단을 잘 못하겠는데… 일단은 전반적으로 읽을 만하니까 4를 주고…

[채점자 C-02]
Target reader는… 그러니까 이건 글이 독자에게 얼마나 적절하냐 이건데... 여기서는 좀 맞지 않는 게 눈에 뜨이네.. 4등급 줘야겠다. 글도 길게 쓰고 성의있게 쓰기는 했는데.. 그래서 좀 더 점수를 높게 주면 좋은데.. 근데.. 뭐 더 이상은 줄 수가 없네..

본 연구에서의 채점자가 평가를 위해 사용한 평가 척도의 내용에 대해 이해의 어려움을 겪지 않도록 1:1의 개별 훈련 과정을 가졌는데도 불구하고 채점자들이 평가 척도 내용의 이해 및 적용에 있어서 위 발췌에서 보듯이 어려움을 겪었던 것으로 나타났는데, 이와 같은 현상은 여러 가지 원인으로 인해 발생할 수 있다. 즉, 척도에 대한 설명이 불충분했기 때문이거나, 척도의 내용이 평가하는 글의 수준들과 부합하지 않아 평가하는 도중 평가 척도에 대한 자신의 이해에 대해서 의심과 의문을 계속해서 갖게 되었기 때문일 수도 있다. 이와 같은 결과로 미루어 볼 때 채점자 훈련 시 채점자들이 사용하게 될 평가 척도에 대한 이해를 충분히 할 수 있도록 돕는 과정이 필요할 것이다.

셋째, 아래의 발췌에서 보듯이 평가하는 글에 대하여 부여 가능한 두 개 정도의 점수 간에서 어떠한 것으로 결정해야 할지를 고민하는 행위에 주목할 필요가 있다.

[채점자 B-06]
(생략)보면은... 내용 측면에서는.. 장점 단점 나누기는 했는데…뭐.. 전반적으로 내용이…충실하게 있는 게 아니라 장점도 있고 단점도 있다고 하고는 설명이 제대로 안 되어 있는 것 같아서…이거는 좀… 그런데 이것도 어쨌든 장단점을 언급을 했으니까 major에서 내용이 빠졌다고 보기는 그러니까 3으로 줘야 될 지 4로 줘야 될지 모르겠는데…일단 4로 줘야 될 것 같고…

(생략)

[채점자 A-06]
(생략) 내용은 아.. 이거 뭐 한 5를 줘도 괜찮고… 어법도.. 아주 뭐.. 비교적 정확한 편이고..5를 줄까..? 5를 줄까…? 5는 너무 지나친 것 같아서... 한 4...? (생략)

위의 발췌에서와 같이 최종 점수를 결정하는 데에 있어서 채점자가 두 점수 사이에서 고민하는 경향들이 많이 발견된다. 이처럼 점수 결정의 어려움을 나타내고 가능한 두 점수 간에 고민하는 현상들로 미루어 볼 때, 채점자 훈련의 과정에서 채점자들에게 이 과정에 대한 훈련이 필요할 것으로 보인다. 즉, 이러한 과정 자체를 제거하도록 한다기 보다는 그러한 과정에서 어떠한 방식으로 결정을 내려야 할지에 대한 훈련이 필요할 것으로 보인다. 뿐만 아니라 평가 척도를 개발할 때에 채점자의 이러한 특성을 고려하여 평가 척도를 개발하는 것 또한 필요할 것이라고 본다. Alderson(1991)에 의하면 평가 척도는 사용자 중심의 평가 척도(user-oriented rating scale), 평가 개발자 중심의 평가 척도(constructor-oriented rating scale), 채점자 중심의 평가 척도(assessor-oriented rating scale) 이렇게 3가지 유형으로 나누어 볼 수 있는데, 채점자들의 이러한 평가 행위상의 특징을 고려하여 이들의 채점 과정을 돕는 채점자 중심의 평가 척도가 되도록 하여 채점자가 평가할 때 의사 결정의 어려움이 최소한으로 되도록 돕는 것이 바람직할 것이다.

넷째로, 채점자들이 점수를 부여하는 과정과 관련하여 주목할 만한 사항은 채점자들이 평가 척도를 사용하여 평가하고 있는데도 불구하고 평가 척도의 기준에서 벗어나지 않기 위해 척도의 내용을 언급해가며 지속적으로 노력하고 있지는 않다는 것이다.

[채점자 B-05]
내용을 보면, 일단은 장점과 단점은 쓰고 있는데…세부사항에 있어서 기준이 명시가 안 되어 있기 때문에 이거를 모든 내용이 충실하다고 봐야 되는 건지… 그래서 6인지.. 아니면 5로 해서, 주요 내용인데 minor가 없는 걸로 봐야 되는 건지.. 그런데, 4에서 some minor가 없다는 거랑… content 측면에서 4하고 5하고 6이 늘 애매한 게 있어서 사실은 거의 늘 주관에 의해서 등급을 분류하는 거랑 차이가 없는 걸로 생각이 드네… 일단은 여기서는… 세부사항에 대해서 설명이… 일단은 장점도 두 개 중에 하나는 길게 설명하고 하나는 짧게 설명했으니까 minor가 완벽하다고 볼 수는 없으니까 5를 줘야 될 것 같고.. (중략) 근데 register, format... 이 항목은... 글쎄.. 이것도 4등급쯤.. 이 좋을 것 같은데... 근데 이거는 뭐를 평가해야 하는 건지 잘 모

르겠는데....(생략)

[채점자 A-06]
다른 학생들보다는 수준이 한 차원 높은 것 같으네… 내용은 아.. 이거 뭐... 한 5를 줘도 괜찮고… 어법도.. 아주 뭐.. 비교적 정확한 편이고..5를 줄까..? 5…..? 5는 너무 지나친 것 같구.... 한 4...? 어휘도.. 필요한 어휘를 잘 썼다고 봐서 5를 줘야겠고.. 글의 구성이나 응집력도 굉장히 높은 것처럼 생각이 되어서 역시 5를 주는 게 좋을 것 같고…글의 어떤 목적에도 거의 뭐.. 한 90% 정도까지 달성하는 것 같으니까 5를 줘야겠고..

채점자가 평가 척도를 사용하지 않고 자신의 주관적 기준에 의하여 평가를 하는 상황이라면 이러한 행위는 문제될 것이 없겠으나 본 연구에서처럼 채점자들에게 평가 척도가 주어져 있는 상황에서는 평가 척도에 가능한 한 최대한 맞추어 평가하는 것이 요구되는 상황이다. 평가 척도에는 평가의 목적, 그 평가에서의 구성, 각 등급의 수준 등이 압축되어 일목요연하게 체계적으로 표기되어 있다. 이것에 맞추어 평가해야 그 평가의 구성 타당도를 유지하는 방향으로 나아갈 수 있게 된다. 그런데 세 명의 채점자들 중 이러한 평가 척도에 고수하려는 면에 있어서 문제가 없는 채점자는 없었다. 이들은 평가 척도를 사용하면서도 이에 고수하기 보다는 채점자의 주관적 기준을 완전히 배제하지 못하고 있음을 보였다. 물론, 채점자의 주관적 기준이 사용하는 평가 척도와 일치하다면 이러한 경향이 문제가 되지는 않겠지만, 평가 척도와 채점자의 주관적 기준은 서로 정확히 일치할 가능성은 매우 희박하다는 것을 고려할 때 이와 같은 방식으로 채점자들이 평가를 한다면 이는 평가의 구성 타당도에 영향을 미치는 중요한 문제가 될 수 있게 되며 채점자 훈련 시 고려해야 할 중요한 측면이라 할 수 있다.

다섯째, 채점자들은 점수를 부여할 때 앞서 평가한 글과 비교하며 점수를 부여하기도 한다는 점이 주목할 만하다.

[채점자 B-05]
(생략) 음... 글의 내용에서 어떤 내용이 major이고 minor인지 기준 정하는 게 쉽진 않은데..... 이 학생도 앞의 학생하고 비슷하니까 뭐... 한... 4쯤....? (생략)

[채점자 C-06]
(생략) 이 학생은... 다른 학생들에 비해서 좀 더 잘 했으니까..... 뭐, 이것저것 문법적인 것도 잘 맞춰서 썼고... 내용도 더 자세하고....
위의 발췌에서도 보듯이, 채점자들은 평가하는 글에 점수를 부여하기 위해

사용하는 평가 척도만을 기준으로 하여 평가하는 것이 아니라 바로 앞서 평가한 글들의 수준, 그리고 그 글들에 자신들이 부여한 점수들을 또 하나의 평가 기준으로 사용하여 평가하고 있었다. 이렇게 될 경우 평가되는 글은 평가 척도를 기준으로 객관적이고도 일관성 있게 채점되지 못할 가능성에 놓일 수 있게 된다. 평가 척도가 사용되는 데에도 불구하고 그 글의 앞에 어떠한 수준의 글이 놓이느냐에 따라 상대적으로 평가될 우려가 있게 되는 데 이것은 평가의 신뢰도 유지를 위협할 수 있게 된다.

마지막으로, 채점자들은 분석적 평가 척도를 사용하며 평가하고 있었는데, 그들이 점수를 부여하는 과정에 있어서 하나의 평가 항목에 대한 평가가 다른 평가 항목들을 채점할 때에 영향을 미치는 후광 효과(halo effect)가 있음을 발견할 수 있었다.

[채점자 A-02]
어휘는.. 뭐 내용과 비슷하게 3 정도가 좋겠고…그 다음에 글의 구성이나 응집성은... 단락은 나누고 있기는 했는데 이게.. 알맞게 나눠진 것이라기보다는 그냥.. 한 두 줄씩 나눈 것에 지나지 않은 상태라 되어서.. 그래도 뭐.. 일단 단락을 나누어서 글을 전개했으니까 한 3 정도가 좋을 것 같고… 그 다음에.. 아… 처음에 편지글에 맞게… 처음에… 좀 부족하기는 하지만 뒷부분에 with love 같은 표현이 있어서 소개하는 글에는 맞으니까 한 3 정도가 좋을 것 같고.. 그 다음에.. 뭐 글의 흥미도도 .. 탠저린이나 옥돔, 돌하루방 이런 등등을 소개한 것으로 봐서 흥미도도 조금 있다고 보여지고 그래서 한 3 정도 주는 게 좋겠다.

일반적으로, 분석적 척도를 사용할 경우 평가하는 글을 여러 측면에서 독립적으로 평가하여 평가 결과를 프로파일의 형태로 제시한다는 것이 본래의 목적인데 반해서, Hill과 Storch(1994)가 언급하듯이 분석적 평가 항목 중 하나의 평가 측면에 대한 평가가 나머지 평가 항목의 평가에 영향을 미치는 현상이 발견되곤 한다. 그런데, 본 연구에서의 채점자들도 이와 같은 현상을 나타내어 보이는 경우들이 있었다. 평가 척도를 사용하여 채점하는 경우가 평가 척도를 사용하지 않고 주관적 평가 기준에 따라 채점할 때보다, 또한 통합적 평가 척도를 사용하는 경우가 분석적 평가 척도를 사용할 때보다 채점 과정에서 더 많은 노력과 시간을 요구하게 된다. 따라서 그러한 분석적 채점 과정에서 나타날 수 있는 후광 효과로 인해서 채점에 부정적 영향을 주는 가능성이 최대한 줄어들 수 있도록 채점자 훈련을 할 필요가 있을 것이다.

V. 결론 및 제언

이상에서 살펴본 바와 같이 쓰기 기술의 평가를 위해서 채점자들이 구체적인 평가 기준을 가지고 채점하게 될 때, 채점 과정 및 결과는 채점자들의 배경에 따라 영향을 받을 수 있게 된다. 따라서 여러 명의 채점자들이 참여하여 채점을 하는 경우 평가의 타당도 및 신뢰도 유지를 위해서 채점자 훈련 과정이 필요하게 된다. 채점자 훈련은 채점에 있어서 단지 채점자들 간의 일치도를 높이기 위함보다는 각 채점자가 채점자 내의 일관성을 유지하는 것이 더 추구해야 할 목표라고 할 수 있을 것이다. 채점자 간의 일치도는 물론이고 채점자 내 일관성의 문제는 단순히 신뢰도상의 문제로 그치는 것이 아니라 더 근본적으로 평가의 타당도에 기반한 문제이므로 간과할 수 없는 중요한 측면이 된다. 따라서 이를 위해서는 채점자들이 채점 과정에서 어떠한 평가 행위를 보이는 지를 알아보는 것이 선행되어야 할 과제로 여겨진다. 따라서 본 연구에서는 현직 영어 교사들로 하여금 이미 개발된 평가 척도 중의 하나인 동시에 한국 고등학생들의 영작문 수준에 가장 근접할 것으로 여겨지는 FCE 영작문 평가의 평가 척도를 사용하여 채점하도록 하고 그 채점 과정 중에 어떠한 평가 행위를 보이는 지를 '소리내어 말하기 과업'이라는 질적 연구 방법을 통하여 살펴보았다.

그 결과, 채점 과정 중 가장 핵심이라 할 수 있는 점수를 부여하는 행위와 관련하여 주요 측면 6 가지를 발견하였다. 채점자들은 평가 자료를 평가하기 전에 그 글에 대한 첫 인상을 중심으로 먼저 반응하는 양상을 보이기도 하였고, 사용하는 평가 척도에 대한 이해의 어려움을 보여준 행위들도 발견되었다. 또한, 평가하는 글에 대하여 부여 가능한 두 개 정도의 점수 간에서 어떠한 것으로 결정해야 할지를 고민하는 행위에 주목 할 만하고 채점자들이 평가 척도를 사용하여 평가하고 있는데도 불구하고 평가 척도의 내용을 언급하며 그 기준에서 벗어나지 않기 위해 노력하는 흔적은 쉽게 찾아 볼 수 없다는 점도 발견되었다. 마지막으로, 채점자들은 분석적 평가 척도를 사용하며 평가하고 있었는데, 그들이 점수를 부여하는 과정에 있어서 하나의 평가 항목에 대한 평가가 다른 평가 항목들을 채점할 때에 영향을 미치는 후광 효과(halo effect)도 보여주고 있었다.

본 연구에 참여한 채점자들에게서 이와 같은 현상들이 공통적으로 발견된 것이 채점자들이 모두 우리나라 중등학교에서의 일반적인 영어 교사라는 직업적 배경에 기인한 것인지 아니면 그 외의 배경을 가진 모든 채점자들에게서 나타나는 현상들인 지는 분명치 않다. 이를 위해서는 향후에 이들과 다른 배경을 가진 채점자들을 대상으로 하여 동일한 실험을 실시한 후 이들 간의 채점 행위를 비교해보는 연구가 필요할 것이다. 하지만, 본 연구에서 얻어진 이와 같은 결과들은 성별이나 근무 경력, 또는 근무하는 학교의 유형에 관계없이 본 연구에 참여한 우리나라 일반적인 영어 교사들에게서 공통적으로 나

타난 현상이므로 앞으로 일반 영어 교사들로 구성된 채점자 집단의 훈련 방안을 계획하는 데에 참고할 수 있는 자료는 될 것으로 여겨진다.

따라서 앞으로 채점자 훈련을 전문적으로 받지 않은 교사들이 수행 평가의 일환으로서 영작문 평가를 해야 할 상황들이 늘어나고 있는 현 시점에서 이들을 위한 효과적인 채점자 훈련 방안 또는 가이드라인 강구를 위해서 다음과 같이 제언하고자 한다.

첫째, 평가 척도를 사용할 경우 우선 평가 척도에 대한 이해를 충분히 할 수 있도록 해야 하고 채점하는 동안 주관적 기준이나 글의 첫 인상보다는 평가 척도에 밀착해서 채점할 수 있도록 해야 할 것이다. 그럼으로써 평가의 타당도를 높이고 그 결과 최소한 채점자 내 신뢰도를 유지할 수 있게 될 것이다. 그러기 위해서는 기본적으로 쓰기 능력에 대한 일반적인 정의에 대한 이해를 먼저 할 수 있어야 할 것이다. 쓰기 능력이라는 것이 단지 문장 수준에서 정확한 문장을 쓰는 것이 아니라 문단과 담화 수준에서 글을 구성하여 자신의 의견과 생각을 독자에게 전달하는 의사소통 활동이라는 것에 대한 이해를 위한 활동 및 안내가 채점자 훈련 안에 있어야 할 것이다. 그 후 평가의 타당도를 고려할 때 채점 과정 중 평가 척도의 내용에 충실히 하는 것의 중요성에 대한 인식을 높일 수 있는 활동도 포함되어야 할 것이다.

둘째, 교사들은 통합적 평가 척도보다 분석적 평가 척도가 L2 화자들의 능력 평가에 긍정적인 피드백을 줄 수 있다는 점에서 분석적 평가 척도의 유용성에 대해 주지하도록 해야 할 것이고, 그와 더불어 분석적 평가 척도를 사용하여 채점하는 동안 평가 항목 중 어느 한 항목에 의해 나머지 항목들에 대한 채점이 영향을 받는 후광 효과가 나타나지 않도록 이에 대한 훈련을 해야 할 필요성이 있겠다. 통합적 평가를 하는 경우보다 분석적 평가를 하게 되면 비용이나 노력, 시간 면에서 더 많은 자원을 요구하게 되는데 이러한 기회비용을 고려할 때, 학습자들에게 분석적 평가를 함으로써 진단적/분석적 피드백을 해줄 수 있다는 효과가 최대한 발휘될 수 있도록 해야 할 것이다.

셋째, 채점자들의 평가 양상을 살펴보면 하나의 샘플에 최종 등급을 부여하기 전에 두 개 정도의 후보 등급을 고려해놓고 그 중 고민하다가 결정하는 경향이 있는 것을 고려할 때, 이들의 이러한 채점 행위를 반영하여 평가 척도를 개발하여 이들의 채점 과정을 도울 수 있을 것으로 보인다. 그러기 위한 평가 척도의 형태는 다양하게 구상해볼 수 있을 것이다. Upshur와 Turner(1995)가 제시한 것처럼 Empirically derived, Binary-choice and Boundary-definition (EBB) 평가 척도[7]의 형태를 취함으로써 어느 자질에 대해 Yes 또는 No로 양분하게 하여 평가의 등급을 정해 나가는 방식도 가능할 것

7 실제 수행 자료를 분석하여 척도를 개발한 예 중의 하나로서, 이 척도는 각 등급의 전반적인 특성보다는 각 등급이 이웃하는 등급과 구분되는 특성이 무엇인지에 초점을 맞추어 이분법적인 방법으로 평가해 나가도록 척도의 내용이 기술되어 있다.

이고, 이와 달리 평가 항목과 등급으로 구성된 일반적인 틀(grid) 형태를 취하되 각 등급에 대해 기술할 때 그 등급의 특성의 전반적인 내용을 기술할 뿐만 아니라 각 등급의 차상위 등급 또는 차하위 등급과 구분되는 자질이 무엇인 지를 기술하는 형태의 척도도 가능할 것이다. 이런 식으로 채점자들의 평가 상의 행위를 바탕으로 개발된다면 채점자가 평가 척도의 사용 과정을 도움으로써 그들이 평가의 일관성 및 타당도를 유지하도록 하는 데에 긍정적인 도움을 줄 수 있을 것이라 여겨진다.

이상에서 고찰한 바와 같이 영작문 평가 시 채점자들은 평가 척도를 사용한다 하더라도 채점에 있어서 여러 가지 한계 및 어려움을 계속해서 가지고 있게 된다. 따라서 채점자 훈련 과정은 이러한 취약점을 극복하기 위한 방향에 초점을 맞추어서 효과적으로 이루어지도록 해야 할 것이다. 이와 같은 채점자 훈련은 특정 채점의 상황에서 특정의 평가 척도에 대해 하는 것도 가능한 반면, 교사 연수 프로그램을 통하여 일반적인 훈련을 하는 방안도 가능하리라고 본다. 이에 본 연구는 전문적인 채점자로서의 훈련 경험이 없는 우리나라 영어교사들이 학습자들의 영작문을 채점해야 하는 상황이 증가되고 있는 현 시점에서 이들을 위한 채점자 훈련 방안 또는 자가 훈련 프로그램 마련을 위한 기본 자료로서 활용될 수 있을 것으로 기대한다.

참고문헌

김형엽. (2006). 영어작문 교육에서 공동협력자로서의 교사의 역할: 교사-학생 면담을 중심으로. *현대영어교육, 7*(2), 96-129.

박경자, 임병빈, 김재원, 유석훈, 이재근, 김성찬, 장영준, 한호. (2001). *응용언어학 사전*. 서울: 경진문화사.

박상옥, 이유진. (2009). 영어교육에서 베껴 쓰기와 요약하여 쓰기의 효과 비교: 사례연구를 중심으로. *현대영어교육, 10*(2), 60-86.

이영식. (2000). 영어작문 평가에 대한 채점자 훈련의 원리. *영어교육, 55*(2), 35-56.

이지연. (2007). 영어 교사들의 주관적 평가 방법에 따른 영작문 평가 시 평가 항목의 타당도. *응용언어학, 23*(1), 147-172.

최연희. (2002). 채점자 훈련이 영어 작문 채점에 미치는 효과에 대한 연구. *응용언어학*, *18*(1), 41-62.

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71-86). London: Macmillan.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples

differently? *TESOL Quarterly*, *25*(4), 587-603.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.

Carrell, P. L. (1995). The effect of writers' personalities and raters' personalities on the holistic evaluation of writing. *Assessing Writing*, *2*(2), 153-190.

Cho, Sookyung. (2009). Patterns of sentence connectors and the effect of instruction on ESL learners' writing. *Modern English Education, 10*(2), 1-22.

Davies, A., & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 795-813). Mahwah, NJ: Lawrence Erlbaum.

Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the test. *Research in the teaching of English*, *15*, 244-255.

Glendinning, E., & Howard, R. (2001). Examining the intangible process: Lotus ScreenCam as an aid to investigating student writing. *Edinburgh Working Papers in Applied Linguistics*, *11*, 42-58.

Gowen, S. (1984). *Writing, rating and personality type*. Paper presented at the ninth annual University System of Georgia Developmental Studies Conference, Athens.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook* (vol. 5). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1990). Second language writing: assessment issues. In B. Kroll (Ed.), *Second language writing: research insights for the classroom* (pp. 69-87). Cambridge: Cambridge University Press.

Hill, K. (1997). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment-Proceedings of LTRC 96* (pp. 275-290). Jyvaskyla: University of Jyvaskyla and University of Tampere.

Hill, K., & Storch, N. (1994). Analytic rating scales: how diagnostic are they? *Melbourne Papers in Language Testing*, *3*(1), 50-65.

Jensen, G. H., & DiTiberio, J. K. (1989). *Personality and the teaching of composition*. Norwood, NJ: Ablex.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Lumley, T. (1995). The judgements of language-trained raters and doctors in a test of English for health professionals. *Melbourne Papers in Language Testing*, *4*(1), 74-98.

O'Loughlin, K. (1992). Do English and ESL teachers rate essays differently? *Melbourne Papers in Language Testing*, *1*(2), 19-44.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3-12.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research*, *10*(2), 157-180.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*, 197-223.

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

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Yi, Jyi-yeon. (2006). *Construction of a rating scale for writing assessment in an EFL context.* Unpublished doctoral dissertation. The University of Edinburgh, Edinburgh, UK.

이지연
총신대학교 영어교육과
156-783 서울특별시 동작구 사당1동 산 31-3
Tel: (02) 3479-0355 / M.P. 010-9983-3559
Email: jyyi@chongshin.ac.kr

Received 2 October 2009
Reviewed 3 November 2009
Revised version received 11 December 2009