영어교육 63권 2호 2008년 여름

EBB를 이용한 중등학교 말하기 평가척도 개발

황 성 삼
(중앙대학교)

Hwang, Sung-Sam. (2008). Constructing rating scales for oral performance tests at secondary schools based on the EBB model. *English Teaching, 63*(2), 307-331.

This article examines the validity of the empirically derived, binary-choice, boundary-definition (EBB) rating scale for oral performance tests at secondary schools. First, this article takes three problems into account of applying a priori rating scales such as ACTFL guidelines in the secondary school context. It then turns to a posteriori data-based EBB rating scale, and shows advantages of using the inductive scale which can cover the three problems mentioned in the deductive rating scales. Subjects are 20 high school students in an English conversation course and three Korean English teachers. Speech data are collected from the pupils who are asked to tell stories about their past volunteering experiences. Two of the three Korean teachers of this research collaborate to construct the EBB rating scale. Findings indicate that the EBB rating scale mirrors classroom learning in the assesment, divides the subjects in half, and its binary algorithms are clear. ICC inter-rater reliability coefficient is .78 for average measures which is a little lower, but still consistent with the previous result by Upshur and Turner (1995). This paper concludes that the EBB rating scale seems to have high degree of validity and is applicable to the secondary schools.

I. 서론

국가 영어능력인증제 발표로 중등학교에서 말하기 평가에 대한 관심이 높아지고 있다. 현재 말하기 시험이 필요한 영어과목은 인문계 고등학교 2·3학년들이 이수하는 영어회화 과목이다. 교육과학기술부(2006)에 따르면 전국 인문계 고등학생의 41%가 영어회화 과목을 이수한다. 중학교 역시 47%가 한 학기에 최소한 1회 말하기 시험을 실시한다(김해선, 1999). 보통 100점 만점 중 적게는 10점, 많게는 20점까지 직접평가 형태로 말하기 시험을 실시한다. 이처럼 말하기에 대한 관심이 높아져, 교사들은 현장에서 당장 쓸 수 있는 말하기 평가척도를 절실히 원하고 있다. 그러나 중등학교에서 즉각 적용 가능하면서도 타당한 말하기 평가척도 개발 논의가 아직 부족한 편이다.

현재 논의되고 있는 말하기 평가척도들은 주로 공인 영어능력 시험의 언어학적 모형들이다. Educational Testing Service(ETS)의 Test of Spoken English(TSE)

경우 의사소통적 언어능력모형(Bachman & Palmer, 1996; Canale, 1983; Canale & Swain, 1980; The Council of Europe, 2001)에 근거하여 기능적 능력, 사회 언어적 능력, 담화능력, 언어능력을 평가한다(ETS, 2001). American Council on the Teaching of Foreign Languages(ACTFL) 능력시험 평가지침서(ACTFL, 1986) 역시 간접적(implicit) 등급 기술표이기 때문에 채점자의 사전 훈련 없이 사용할 수 없다(Lowe, 1985).

이들 공인 말하기 시험들은 보편적 언어평가 모형으로 중등학교의 교육과정을 채점 기술표에 반영하지 않아(Fulcher, 2007) 수업과 평가를 연계시키기 어렵다. 결과적으로 공인 말하기 시험은 교육과정 이수여부 측정이라는 학교평가의 본래 목적(이완기, 2003)을 잘 달성하기 어렵다. 또한 공인 말하기 시험은 원어민 전문가의 직관과 경험에 근거한 연역적(a priori) 평가척도(Fulcher, 1996, 2007)로 한국 중등학교 학생들의 발화경향을 잘 반영하기 힘들다. 따라서 많은 교사들은 자신들의 경험과 직관에 근거하여 교실상황에 맞는 맞춤형(local) 평가척도(Upshur & Turner, 1995)를 직접 만든다. 그러나 교사들 자신이 만든 말하기 평가척도 역시 공인 말하기 능력시험의 평가척도를 크게 벗어나지 못한다. 원어민의 높은 발음기준과 모호한 문법, 어휘기준을 평가 기술표에 제시하기 때문에 학생은 물론 교사 자신도 채점 신뢰도에 만족하지 못한다(이인제 외 13인, 2004).

이에 관해 최인철(2005)은 전국 중등 영어교사 82명을 대상으로 수행평가 관련 6점 Likert 척도 설문조사를 실시했다. 먼저 말하기 평가의 신뢰도에 대해 평균 3.53(표준편차 .69)로 교사들은 다소 반신반의했다. 반면, 김해선(1999)에 따르면 학생과 학부모는 교사의 채점을 신뢰하는 것으로 나타나, 말하기 시험의 신뢰도 향상을 위한 환경은 어느 정도 마련된 것으로 보인다(오준일, 2006; 이인제 외 13인, 2004). 한편, 최인철(2005)에 의하면 교사들은 말하기 평가를 원어민 교사가 실시하는 것이 도움이 된다고 평균 4.29(표준편차 .57)로 대답해 원어민 의존도를 보였다. 그리고 '말하기 시험이 학생들의 의사소통능력 신장에 도움이 되는가?'라는 질문에 평균 3.56(표준편차 .74)로 다소 불확실하게 긍정적 환류 가능성을 인정했다.

이러한 중등학교 말하기 평가척도의 빈곤과 채점에 대한 자신감 부족은 타당도가 부족한 말하기 평가나 평가 회피로 이어질 수 있다. 교실평가의 특수성을 반영하면서도 신뢰성이 있고 타당한 구어 평가척도가 필요하다. 따라서 본 연구는 다음 장에서 연역적 말하기 평가척도를 교실평가에 적용시 발생하는 문제점들을 먼저 검토해 본다. 그리고 수험자의 발화에 근거한 귀납적(a posteriori) 말하기 평가척도인 Upshur와 Turner(1995)의 Empirically derived, Binary-choice, Boundary–definition(EBB) 모형을 소개하고, 중등학교 말하기 수행평가에 EBB 평가척도를 적용해 보고 그 타당성을 탐색해 보고자 한다.

II. 이론적 배경

1. 교실평가와 연역적 말하기 평가척도

ACTFL 지침서류의 연역적 말하기 평가척도는 보통 고부담의 큰 시험에서 많이 사용된다. 높은 채점 신뢰도를 바탕으로 복수의 전문 채점자가 신속하게 대량 채점을 할 수 있기 때문이다. 그러나 연역적 평가척도는 다음의 3가지 측면에서 교실평가와 같은 작은 시험상황에서 적용이 어려워 보인다.

첫째, 연역적 말하기 평가척도는 학생들의 발화경향을 제대로 예측하기 힘들다(Fulcher, 1987). 이와 관련, Shohamy(1995)는 언어능력은 실제 발화표본(sample of performance)을 통해서만 나타난다고 했다. 평가척도와 수험자 발화경향과의 일치는 좋은 평가척도의 기본요건이다(Upshur & Turner, 1995). 연역적 평가척도는 수험자가 어떻게 반응할 것이라고 연구자가 이론에 근거하여 예측하지만 실제 나타나지 않는 경우가 많다(Fulcher, 1987). 발화경향과 평가척도가 불일치되면 수업내용과 평가가 단절될 수 있다. 이러한 수업과 평가의 단절은 학생들에게 시험공포, 불공정성, 무기력감, 정의감 결여, 포기, 편견, 의혹, 실패감을 초래하여 시험에 대한 부정적 환류를 일으킬 수 있다(Shohamy, 2001). 연역적 평가척도를 교실평가에서 쓰기 위해서는 학생들의 발화은행이 구축되고(Fulcher, 1987) 발화경향이 분석되며, 하위등급의 세분화가 선행되어야 한다. 이와 관련, 일본은 평가와 학교교육과의 연계성을 높이기 위해 ACTFL의 중급단계를 3단계에서 5단계로 세분화시킨 Standard Speaking Test(SST)를 개발했다(ALC, 2007, 오준일, 2006 재인용).

둘째, 연역적 말하기 평가척도는 대체로 원어민 중심의 상대평가이다(신동일, 2002; 신동일, 서인영 2002; Bachman & Savignon, 1986; Lantolf & Frawley, 1985). Foreign Service Institute(FSI) 전통을 따르는 ACTFL, ETS, Interagency Language Roundtable(ILR)의 말하기 평가척도가 '교육받은 원어민'의 언어능력을 최상위 절대등급으로 놓고 나머지 등급을 상대적으로 정의하기 때문에(Wilds, 1975) 상대비교적 평가척도가 된다. Fulcher(2003) 역시 FSI 전통의 말하기 평가척도들이 '외국어 사용능력이 없는 사람(no proficiency)'의 최하위 등급에서 '교육받은 원어민(well-educated native speaker)'의 최상위 등급까지 위계적으로 등급을 정하고 상대 비교우위적 기술방법으로 평가한다고 비판한다. Carroll(1982)에 의하면 영국 English Language Testing Service(ELTS)[1]의 말하기 평가 또한 평가 기술표에 근거해 절대평가라고 주장하지만 실제로는 상대비교의 상대평가 족쇄에 묶여 있다고 지적한다.

이에 대해 신동일(2001)은 말하기 절대평가 시험들이 목표지향적(criteria-referenced) 성격을 충실히 반영 못하는 이유는 준거적 기술표 안에 원어민 표준의 평가

[1] ELTS는 현재 International English Language Testing Service (IELTS)로 바뀌었음.

관행이 너무 당연시되기 때문이라고 지적한다. Skehan(1984) 역시 등급 간에 독립적인 준거 제시보다는 상대 비교 우위적인 기술방법이 보편적으로 사용되기 때문이라며 등급기술에 문제를 제기한다. 그래서 교사들이 상대비교평가 타입의 말하기 시험 평가척도를 바로 사용하기가 어렵다(신동일, 서인영, 2002). 말하기 평가척도가 너무 모호하고(Upshur & Turner, 1995) 연역적이기 때문에(Fulcher, 1996) 평가기술표 내용 중 일부만이 사용되거나, 원어민이 없는 학교에서는 말하기 시험 자체를 포기하는 원인이 되기도 한다.

이러한 원어민 중심의 FSI 말하기 평가 전통(Adams & Frith, 1979)은 말하기 시험문제를 넘어 국가적 손실로 이어질 수 있다. 전병만, 박준언, 유제명, 최희경(2006)에 따르면 정부의 영어교육 관련 총예산 680억원 중 50%인 340억이 원어민 교사 채용으로 들어간다. 그런데 전국에서 1명이상 원어민 교사가 파견된 초·중등학교의 비율은 평균 18.4%에 그친다. 전체 예산의 50%를 쓰면서도 원어민 배치율은 18.4%에 그쳐, 예산집행의 효율성이 떨어져 보인다. 학교 급별 원어민 배치현황은 초등학교 14.2%, 중학교 21.8%, 고등학교 19.2%이다. 정부의 한정된 영어교육 예산을 원어민 교사에만 다 투여할 수 없는 상황에서, 원어민 중심의 말하기 평가 전통은 개선의 여지가 크다.

한편, Common European Framework of References(CEFR) 척도의 경우 '교육받은 원어민' 기준을 빼고 각 등급의 기술내용을 '할 수 있다(can-do)' 식으로 기술하여 원어민 중심의 평가관행을 탈피하고자 했다(Hudson, 2005).

셋째, 연역적 말하기 평가척도는 교실평가의 특수성을 잘 반영하기 힘들다. 중등학교의 말하기 평가에서는 학생간의 차이가 작고, 평가척도에 대한 채점자 훈련도 어렵다(Upshur & Turner, 1995). 따라서 ACTFL Oral Proficiency Interview (OPI)와 같은 공인 말하기 시험에서 다수의 중등학생들이 특정 세부등급에 몰려 평가 변별 자체가 어려울 가능성이 높다. 중등학생들의 등급예측과 관련하여 고등학교 1학년 학생의 말하기 성취기준을 김덕기, 안병규, 오윤자, 김영규(1999)는 ACTFL-OPI 중급-상, 박기화(2005)는 중급-중, 이병민(2003)은 중급-하 수준으로 설정했다. 이러한 말하기 등급예측을 이병민(2003)은 교육과정 연구로, 박기화(2005)는 문헌연구로, 김덕기 외 3인(1999)은 8명의 고등학생 말하기시험 결과분석으로 목표등급을 설정했다. 그런데 이러한 연역적 평가척도의 등급 기술표가 다양한 교실 수업목표와 부합되는 경우가 드물기 때문에(Upshur & Turner, 1995) 탈맥락적 평가척도(Hudson, 2005, Fulcher, 2007 재인용)가 되기 쉽다.

실증적으로 확인되지 않은 말하기 능력 등급예측의 경우 실제 교실평가에서 척도의 타당성을 보장받기 힘들다. 평가자의 직관과 경험에만 의존한 연역적 평가척도가 자칫 심리적 현실(psychological reality)이 되기 쉽기 때문이다(Fulcher, 2003). 이런 이유로 Liskin-Gasparro(1996)는 ACTFL-OPI 시험의 발화표본을 담화분석하여 등급 기술표의 정확성과 위계성을 검증하자고 제안했다. 결론적으로 교실 말하기 평가에서 연역적 말하기 평가척도는 척도와 발화표본과의 불일치, 원어민 중심의 상대평가, 교실평가의 특수성의 이유로 현장적용에

난항이 예상된다.

2. 말하기 표본에 근거한 EBB 평가척도

1990년대 들어와 경험과 직관에 근거한 말하기 시험이 타당성 결여의 가능성을 안고 있다는 문제점이 심각하게 지적되어(오준일, 2006) 말하기 표본에 근거한 평가척도가 제안(신동일, 2002; Fulcher, 1987, 1988; Pienemann, Johnson & Brindley, 1988; Shohamy, 1990)되고, 평가척도 개발이 시작된다(Fulcher, 1996; North, 1994; Upshur & Turner, 1995, 1999). Matthews(1990) 역시 평가가 취할 수 있는 가장 현실적 접근법은 특정 모집단의 수험자를 대상으로 좁게 평가하는 것이라 주장한다. 그래서 수험자가 어떻게 반응해야 한다는 규범이 아닌, 관찰된 학습자의 실제 발화행위에 근거(Pienemann et al., 1988)한 실증적 평가척도 개발(Fulcher, 2003)이 필요하다.

기존의 연역적 평가척도의 한계를 극복하고자 Upshur와 Turner(1995)는 말하기 표본에 근거한 이분적 등급판정 평가척도인 EBB 척도를 개발했다. EBB 평가척도는 평가자 중심이고, 전체적(holistic) 채점방식이며, 실제 발화표본에서 평가의 구인을 찾는다(Fulcher, 2003). Upshur와 Turner(1995)는 캐나다 퀘벡주 Montreal시에서 프랑스어가 모국어인 초등학생 101명을 대상으로 영어 말하기 평가를 실시했다. 평가과업은 학생들이 ESL 수업시간에 2분 30초 분량의 'Arnold of the Ducks'라는 애니메이션 영화를 시청한 후, 영화내용에 대해 발표하는 말하기 활동(story retell)이었다.

EBB 평가척도 개발단계에는 총 6명의 연구자가 참여했으며, 이들은 논문저자 2명과 응용 언어학 및 ESL 전공 대학원생 4명이었다. 평가척도 개발팀 6명은 그림 1의 의사소통능력 평가 EBB척도와 그림 2의 문법적 정확성 평가 EBB 척도를 공동 개발했다. 채점단계에서는 논문저자 1명, 대학원생 1명이 각각 한 팀이 되어 첫 번째 채점팀은 의사소통능력 평가 EBB척도로, 두 번째 채점팀은 문법적 정확성 평가 EBB 척도로 각각 채점했다. 채점 결과 의사소통능력 평가 부문에서 .81, 문법적 정확성 평가 부문에서 .87라는 높은 신뢰도를 얻었다.

EBB 평가척도 개발에서 가장 중요한 단계는 전체 수험자들을 좌우로 균등하게 이분하는 등급기준 질문(criterial questions)을 중심으로 알고리즘 그림(decision tree)을 그리는 일이다(Fulcher, 2007). 그림 1의 의사소통능력 평가 EBB 척도에서 좌우 양분하는 등급기준 질문은 'Coherent story retell vs. listing'이다. 채점자가 수험자의 말하기 내용이 단순한 사실의 나열이 아닌 일관성 있는 이야기 말하기로 볼 수 있으면 오른쪽의 'YES'로, 볼 수 없다면 왼쪽의 'NO'로 내려간다. 만약 수험자의 말하기가 일관성이 없어 'NO'로 내려가고 다음 등급기준 질문인 'One story element only or 'garbles' '에서 'YES'로 판단되면 다음 단계의 질문으로 내려가지 않고 최저 등급수준인 1점이 된다. 전체 등급기준 질문

인 ‘Coherent story retell vs. listing’을 통해 1-3등급이 99명, 4-6등급이 99명으로 양분되어 전체 수험자들이 정확히 50%씩 양분되었다. 그리고 채점할 수 없는 결측값은 4명이었다. 채점대상 인원이 총 101명이 아닌 202명이 된 이유는 2명으로 이루어진 채점팀이 개별적으로 101명에게 점수를 부여했기 때문이다.

그림 1
의사소통능력 평가 EBB 척도 (Upshur & Turner, 1995, p. 8)

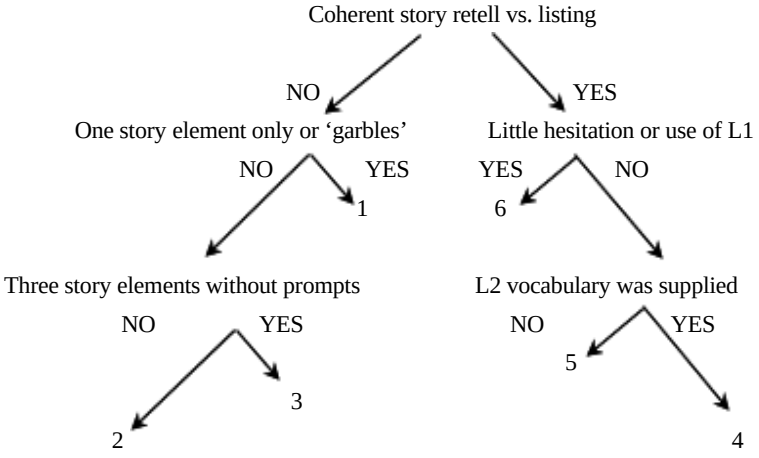
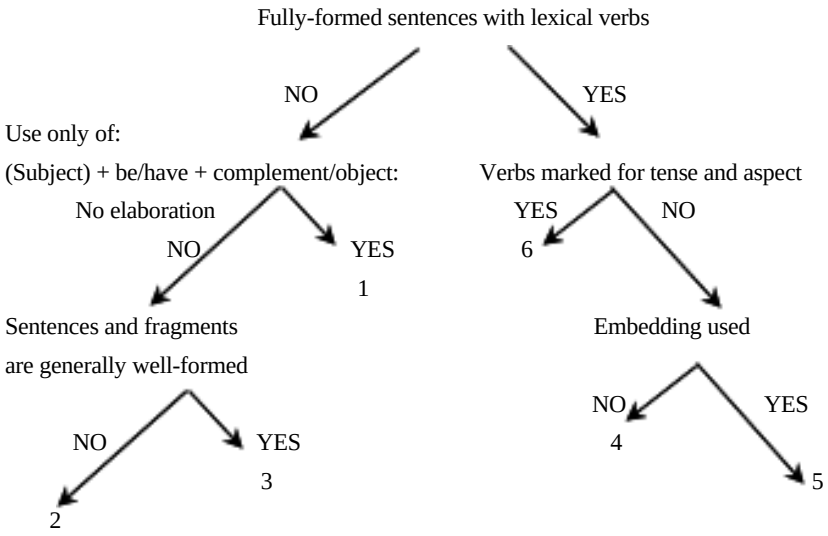


그림 2의 문법적 정확성 평가 EBB척도에는 ‘Fully-formed sentences with lexical verbs’가 전체 수험자를 양분하는 중앙 등급기준 질문이다. 수험자의 발화문장이 완전한 문장으로 이루어져 있다고 채점자가 판단하면 오른쪽의 ‘YES’로, 그렇지 않다면 왼쪽의 ‘NO’로 내려간다. 만약 수험자의 발화가 완전한 문장으로 이루어져 있지 않다면 ‘NO’로 내려가고, 다음 등급기준 질문인 ‘Use only of: (Subject) + be/have + complement/object: No elaboration’에 대해 ‘NO’로 판단되면 다시 ‘Sentences and fragments are generally well-formed’로 내려가고, 여기에서 ‘YES’로 판단되면 3점으로 판정된다.

전체 등급기준 질문인 ‘Fully-formed sentences with lexical verbs’으로 1-3등급이 92명, 4-6등급이 106명으로 양분되어 전체 수험자들이 50%에 근접하게 좌우 양분되었다. EBB 평가척도의 개발과정은 제3장의 연구 절차에서 자세히 소개하고자 한다.

그림 2
문법적 정확성 평가 EBB 척도 (Upshur & Turner, 1995, p. 9)



한편, EBB 척도는 기존의 연역적 평가척도와 다른 언어평가 인식론에서 출발한다. FSI 전통의 기존 연역적 평가척도는 척도 기술표의 모호성을 수용하고(Fulcher, 2007) 구인과 평가방법을 분리하여 탈맥락적으로 언어능력을 평가(Bachman, 1988)하여 절대등급을 부여한다. 반면, 귀납적 평가척도인 EBB는 내용중심, 과업의존적, 맥락적(contextualized) 평가로 구체적인 발화경향 기술이 필요하며 가변적 등급을 부여한다. 고부담의 큰 시험에서는 높은 채점 신뢰도가 요구되는 반면, 저부담의 작은 시험에서는 수업과 평가의 결합(이완기, 2003)을 통한 교육적 환류가 요구된다. 신동일(2002) 역시 평가자 집단이 자체적인 전문성과 경험을 축적하고 있고, 커다란 의사결정을 요구하지 않는 작은 시험이라면 EBB 평가척도를 고려해 볼 만한 가치가 있다고 주장했다. 따라서 저부담의 작은 시험에 적합한 EBB 평가척도는 다음의 4가지 측면에서 현장적용 가능성이 커 보인다.

첫째, EBB 척도는 학교에서 수업과 평가를 연계시키는(Upshur & Turner, 1995) 교실평가에 적합하다. 수업목표와 평가목표가 연동되면 수업내용이 자연스럽게 학생의 발화에 반영된다. EBB 척도 말하기 평가가 학습과 연관지어 평가하기 때문에 통제가 아닌 피드백으로 긍정적 환류를 일으켜(최인철, 2005; Finch & Shin, 2005; Hudson, 2005; Khattri, Reeve & Kane, 1998; Shohamy, 1995) 교수학습이 강화된다.

그리고 교육과정과 평가의 연계는 EBB 평가척도의 내용타당도를 높여 준다. EBB 척도는 문법, 발음, 어휘의 언어능력에서부터 담화, 화용능력에 이르기까

지 교실수업 내용과 관련된 다양한 의사소통능력 단면들을 등급결정 질문으로 만들 수 있다. 그러나 채점 기술표의 선형적 격자에 나타나는 모든 채점국면을 충족해야 하는 연역적 척도와 달리, EBB 척도는 필요한 언어능력 국면들만을 선택적으로 기술할 수 있다.

둘째, EBB 평가척도는 이분적 의사결정으로서 사용하기에 쉽다(Fulcher, 2003, 2007). 이분적 의사결정은 등급구분 질문에 YES/NO로 대답하는 이분적 알고리즘이다(North & Schneider, 1998; Upshur & Turner, 1995). 그래서 등급구분 질문에 등급간 발화차이가 잘 드러나면, 평가자의 배경지식이 적어도 EBB 척도를 사용할 수 있다. 이런 EBB 척도의 실용성은 평가 비전문가도 언어평가, 교육심리측정, 통계에 대한 사전지식없이 시험을 만들 수 있어야 한다는 Davidson과 Lynch(2002)의 주장과도 일맥상통한다. EBB 척도는 한 문장의 간결한 등급질문이지만, 실제 발화에 나타난 언어능력을 평가하기 때문에 수험자를 쉽게 이분할 수 있다. 이런 의미에서 EBB 등급기술문은 기존의 복잡한 등급 기술표보다 등급판정이 용이하다. 사실 “등급표를 자세히 기술해야 꼭 좋은 시험인가?”(p. 208)라고 반문하는 Fulcher(1996)의 주장처럼 자세한 등급 기술표가 항상 타당한 등급판정을 보장하지는 않는다.

셋째, EBB 평가척도는 비교적 명확한 등급경계선을 제시할 수 있다(Upshur & Turner, 1995). 연역적 평가척도가 두 등급 사이의 중간점을 나타내 양 등급의 유사성을 갖고 있어 모호한 반면, EBB 척도는 두 등급의 차이점을 보여줘 등급 사이에 명확한 경계선을 그을 수 있다(Fulcher, 2003; Upshur & Turner, 1995). 그래서 원어민 중심의 상대기술적 등급표를 벗어나 수험자의 세부능력 변별이 가능하고 외국어 습득 이론의 선형적 가정에서 비교적 자유롭다. EBB 척도의 정확성과 관련, 이호(2007, 12월)는 영어교육학과 4학년 학생 34명에게 EBB 평가척도로 동료학생의 말하기를 채점하게 했다. 채점 후 실시한 EBB 척도에 대한 설문조사 결과, 예비교원 학생 중 67%가 ‘내 말하기에 대한 동료 평가가 정확했다’라고 답변했다.

넷째, EBB 평가척도는 안정된 채점 신뢰성을 갖고 있다. Upshur와 Turner(1995)의 EBB척도 개발 연구에서 의사소통능력 평가 부문에서 .81, 문법적 정확성 평가 부문에서 .87라는 높은 신뢰도를 확보했다. 한편, FSI전통의 연역적 평가척도인 ACTFL-OPI에서는 채점자간 신뢰도가 .90이상으로 전반적으로 EBB 채점 신뢰도보다 높게 보고되었다. 예를 들어 프랑스어 ACTFL-OPI에서 전문 채점자와 채점자 훈련 연수생간의 신뢰도를 측정한 Magnan(1987)의 연구에서 채점자간 신뢰도가 .94로 나왔다. 이처럼 ACTFL-OPI의 신뢰도가 EBB 척도보다 상대적으로 높은 이유는 10등급의 넓은 등급간격, 10분 이상의 많은 발화량, 표준화된 채점자, 많은 수험자라는 큰 시험의 평가상황 때문으로 보인다. 반면, Upshur와 Turner(1995)의 EBB 채점 신뢰도는 1분 남짓의 짧은 발화와 개인차가 적은 ESL 교실상황, 채점자 훈련 부재의 평가상황(Upshur & Turner, 1995)에서 도출되었다. 따라서 신뢰도 저해 변수가 많은 교실평가에서

나온 .80이상의 채점 신뢰도는 EBB 척도의 안정성을 잘 보여 준다. 결론적으로 EBB 평가척도는 수업과 평가의 연계성, 이분적 의사결정의 용이성, 등급판정의 명확성, 척도의 신뢰성으로 인해 중등학교에서 그 적용가능성이 높아 보인다.

3. 연구 문제

본 연구에서는 다음과 같이 4가지 연구문제를 설정했다.

1) EBB 평가척도에 교수학습 내용이 반영되는가?
2) EBB 평가척도는 수험자를 좌우 이분하는가?
3) EBB 평가척도의 이분적 등급판정은 명확한가?
4) EBB 평가척도는 안정된 채점자간 신뢰도를 나타내는가?

III. 연구 방법

1. 연구 대상

본 연구대상 학생들은 서울시내에 위치한 인문계 고등학교 2학년 남학생들이었다. 학생들의 나이는 17세에서 18세인 같은 반 33명 중 20명이며, 학생들의 이름은 모두 가명 처리하였다. 총 33명 중 녹음장치의 오작동과 녹음거부 학생을 뺀 20명이 실제 참여했으며, 학급의 영어회화 반 평균은 51.5 표준편차 21.6로 평균적인 인문고교 2학년 이과 남학생 집단이었다. 표 1에 의하면 연구 참여 교사는 모두 3명의 남교사이며, 교사 A, B는 척도개발 및 채점에, 교사C는 채점에만 참여했다. 특히 교사A는 연구자이며 척도개발 및 채점에 참여했다. 한편, 연구참여 대상자들이 본 연구를 위한 한시적 집단이 아니라, 원래 영어회화 수업에 참여한 자연집단이라는 점에서 본 EBB 연구는 생태적 타당성(ecological validity)을 갖는다(신상근, 2007; Brewer, 2000).

표 1
연구 참여교사

	교사A	교사B	교사C
나이	39세	57세	40세
학력	석사	학사	학사
교직경력	만9년	만27년	만7년
담당과목	영어회화	영어회화	영어II
연구참여	척도개발 및 채점	척도개발 및 채점	채점

2. 연구 절차

교사A는 말하기 평가의 주제를 학기 초인 3월에 예고하고 6월 초에 1학기 기말고사 말하기 수행평가로 실시했다. 말하기 주제는 N출판사 영어회화 교과서 7과의 'I'm going to do volunteer work at a foodbank'이다. 말하기 시험주제를 봉사활동 내용으로 정한 이유는 대학입시의 비교과 영역으로 학생들이 봉사활동 경험이 많았기 때문이다. 말하기 시험 과업은 이야기 말하기(story telling)였다. 학생들은 교과서 7과를 먼저 배우고, 실제 또는 가상 봉사활동 경험을 이야기 구성문법(story grammar)을 이용하여 말하는 방법을 배웠다. 이야기 구성문법은 Stein과 Glenn(1979)의 '배경(settings) ⇒ 도입사건(initiating events) ⇒ 내적반응(internal response) ⇒ 시도(attempts) ⇒ 직접적 결과(direct consequences)'를 단순화시켜 만든 '배경⇒ 사건⇒ 결말' 모형이었다. 또한 담당교사는 이야기 말하기에서 과거시제가 필요하고 육하원칙이 나타나야 구체적 묘사라고 지도했다.

1단계로 학생들의 1학기 중간고사 영어회화 성적을 참고하여 대표 말하기 표본 8개를 표집했다. 말하기 대표 표본은 부록 2와 같이 영어회화 지필고사 점수와 대학수학능력 모의 듣기평가 점수로 선정했다. 2단계로 교사의 전문적 판단으로 EBB 평가척도를 개발했다. 먼저 교사 A, B가 서로 만나 자신이 무슨 기준으로 8명의 샘플학생들을 4명씩 상위, 하위로 나누었는지를 서로 설명하고, 상위 4명, 하위 4명을 이분하는 등급기준 질문을 함께 만들었다. 3단계로 교사 A, B가 각기 독립적으로 상위 4명을 6점, 5점, 4점으로 나누었다. 4개의 샘플을 3등급으로 나누기 때문에 한 등급에 1-2명의 학생이 들어갔다. 4단계로 채점자 A, B가 다시 만나 상위 4명을 6점, 5점, 4점으로 나눈 기준에 대해 서로 논의했다. 그리고 상위 4명에 대한 등급구분 질문을 결정했다. 5단계로 상위 4명을 구분한 순서대로 다시 개별적으로 하위 4명을 3점, 2점, 1점으로 구분하고 교사 A, B가 다시 만나 하위 등급결정 질문을 완성했다.

6단계로 완성된 EBB 평가척도를 사용하여 교사 A, B가 학생 20명을 개별적으로 채점했다. 마지막으로 단일 모집단 내에서의 EBB 평가척도 일반화 가능성을 보기 위해 척도개발에 참여하지 않은 교사C가 EBB척도로 동일한 20명을 채점했다. 본 연구는 Upshur와 Turner(1995)처럼 의사소통능력과 문법능력 평가척도로 나누지 않고 하나의 전체적 평가척도를 만들었다. 1분 남짓의 짧은 발화의 말하기 표본으로 문법과 의사소통 영역으로 두 개의 EBB 평가척도를 만드는 것이 무리였기 때문이었다.

3. 분석방법

본 연구는 정성적, 정량적 연구방법을 모두 사용하여 EBB 평가척도를 타당화(validation)시켰다. 먼저 MP3로 녹취한 발화내용을 Atkinson과 Heritage (1984, Lazaraton, 2002 재인용)의 전사규정을 사용하여 학생들의 발화를 담화분석했다.

교수학습 내용이 EBB 척도에 실제로 드러나는 지를 살피고, 등급구분 질문이 수험자를 양분하는 지도 확인했다. 그리고 EBB척도의 이분적 등급판정의 명확성을 알아보기 위해 말하기 표본과 척도와의 대조 분석을 하고 채점교사 및 학생과 회상 인터뷰(think-aloud protocol)를 했다. 따라서 교육과정과 평가의 연계성, 이분적 등급판정의 명확성을 말하기 표본의 담화분석을 통해 정성적으로 검증해 보았다.

다음, 말하기 채점결과를 SPSS 12.1v 프로그램을 이용하여 정량적으로 EBB 평가척도의 신뢰성을 검증했다. 교사 A, B, C의 채점결과에 대한 기술통계를 제시하고, EBB척도가 전체 수험자들을 실제로 이분하는 지를 빈도분석과 누가백분율을 통해 확인했다. 그리고 급내 상관계수(Intra-class Correlation: ICC)를 이용하여 EBB 평가척도의 채점자간 신뢰도를 알아 보았다.

본 연구에서 ICC 상관계수를 사용한 이유는 수험자의 말하기 점수가 등간척도 자료이고, 세 명의 채점자가 참여하는 상황이었기 때문이다. ICC는 일반적으로 데이터가 등간척도이고 3명 이상의 채점자나 3번 이상의 시험을 칠 때, 분산분석을 통해 신뢰구간에서 모수의 불편 추정치를 제공한다(이호, 2005; Shrout & Fleiss, 1979). 본 연구의 ICC 모형에서 수험자는 무선추출이고 채점자는 비무선(fixed)의 이원 혼합모형(two-way mixed model)이며 일관성 모형이다. 개별채점자(single measures)의 ICC 상관계수는 Pearson 상관계수들의 평균치와 비슷하고, 복수채점자(average measures)의 ICC 상관계수는 Cronbach alpha와 동일하다(Hwang, 2007).

4. 연구의 제한점

본 연구에서 말하기 시험은 즉흥적으로 실시되지 않고, 3개월 전 담당교사가 주제와 평가방법을 미리 예고했다. 말하기는 즉흥성이 매우 중요하지만, 중등학교 상황에서 주제가 즉흥적으로 주어질 때, 대다수 학생들의 발화는 서너 문장을 넘기지 못한다. 그래서 담당교사는 말하기 주제를 미리 발표하고, 학생들에게 이야기 원고를 미리 작성하게 했다. 학생들의 발화량은 1분 내외였다.

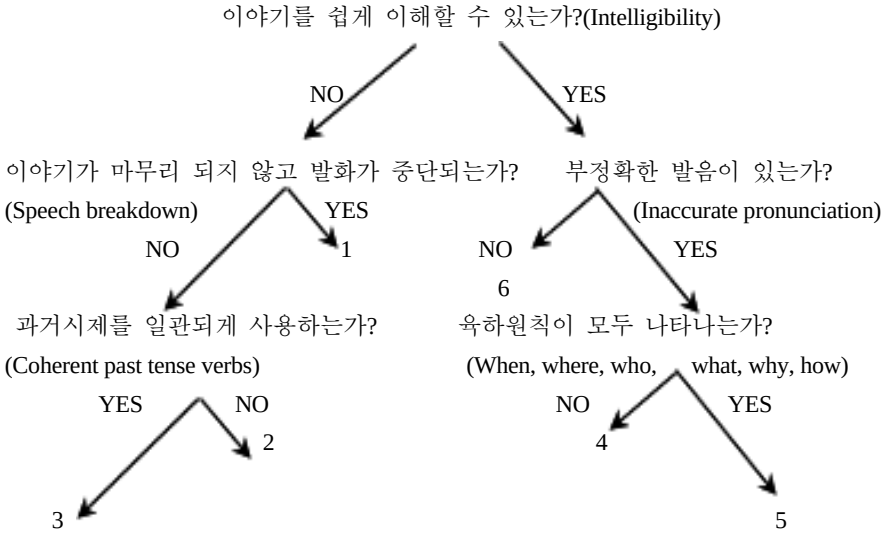
실제 수행평가는 두 가지 방식으로 실시되었는데, 혼자서 이야기 발표하기와 두 사람이 대화하기였다. 녹취하여 평가하는 본 연구의 특성상 대화는 비언어적 요소가 많이 개입되는 관계로 제외했다. 대신, 이야기 발표하기만을 연구대상으로 삼았다. 본 연구에 참여한 학생 20명은 다른 반 학생들과의 형평성에 따라 동일한 형태의 분리평가로 수행평가 점수를 받았다. 분리평가는 문법, 어휘, 발음, 상호작용, 원고의 다섯 가지 영역으로 10점 만점이었다.

IV. 결과 분석

1. EBB 평가척도에 반영된 교수학습 내용

EBB 평가척도 개발 결과 그림 3과 같이 척도가 만들어졌다. 최상위 등급인 6점 등급구분 질문은 ‘부정확한 발음이 있는가?’이며 5점 등급과 4점 등급을 구분하는 등급구분 질문은 ‘이야기의 육하원칙이 모두 나타나는가?’였다. ‘이야기를 쉽게 이해할 수 있는가?’는 전체 상위, 하위등급을 이분했다. 3점과 2점을 결정한 등급구분 질문은 ‘과거시제를 일관되게 사용하는가?’이었고, 최하등급인 1점에 해당하는 등급구분 질문은 ‘이야기가 마무리 되지 않고 발화가 중단되는가?’였다.

그림 3
EBB 평가척도



위에 나온 EBB 평가척도의 등급기준 질문을 통해 교수학습 내용이 채점척도에 반영되었는지를 검증하고자 한다. 먼저 최하위 등급결정 질문인 ‘이야기가 마무리가 되지 않고 발화가 중단되는가?’는 학생들이 수업시간에 배운 이야기 구성 문법과 관련이 깊다. 담당교사A는 이야기 구성 문법을 ‘배경⇒ 사건⇒ 결말’로 설명했다. 즉 배경, 사건, 결말로 이야기를 해야 하나의 완결된 이야기로 볼 수 있다고 수업했다. 2-3 등급의 등급결정 질문인 ‘과거시제를 일관되게 사용하는가?’는 이야기 말하기의 시제를 이해하는 능력이다. 수업 중에 학생들은 이야기 말하기의 과거시제 사용을 학습했다. 수험자를 양분하는 ‘이

야기를 쉽게 이해할 수 있는가?’의 등급구분 질문은 일반적 의사소통 언어능력을 측정한다. 청자가 이야기를 쉽게 이해할 수 있기 위해서 화자가 가급적 완전한 문장으로 응집성과 일관성을 갖고 말하도록 하는 수업이 이루어졌다.

말하기 4-5 등급을 결정한 ‘이야기의 육하원칙이 모두 나타나는가?’라는 이야기 말하기의 묘사능력을 측정한 것이다. 수업 중 학생들은 교사의 이야기 예제를 들으며 ‘주인공이 언제 어디서 무엇을 어떻게 했는가?’라는 육하원칙을 통한 묘사능력을 학습했다. 최고 등급구분 질문인 ‘부정확한 발음이 있는가?’는 점수가 최고 등급으로 몰리는 천정효과를 막기 위해 척도개발 교사들이 정확성의 평가기준을 설정하였다. 그런데 수험자간의 근소한 말하기 능력 차이, 1분 남짓의 말하기에서 최상위 6등급과 4-5 등급 변별해야 하는 부담 때문에 담당교사들이 발음이라는 기존의 원어민 규준에 의존한 것으로 보인다. 그래서 ‘나의 봉사활동 경험’이라는 수업 목표와 정확한 발음이라는 평가기준이 직접적인 관련성을 보이진 않는다.

2. EBB 평가척도의 수험자 이분능력

먼저 EBB 평가척도의 수험자 이분능력을 살펴보았다. 표 2를 보면 20명 학생에게 준 세 교사의 등급 합산 점수가 나와 있다. EBB 평가척도가 총 6점 만점으로 이론적 최고점수는 세 교사의 합산 점수인 18점이다. 그러나 교사 A, B, C가 조금씩 다른 점수를 주어, 실제 최고점수는 16점이다. 같은 방법으로 이론적 최저점수는 3점이고, 실제 최저점수도 3점이다. 학생들의 점수는 점수이면서 동시에 말하기 등급의 성격도 갖는다. 합산점수의 분포에서 모든 학생들을 이분하는 점수는 11점이며, 누가 백분율은 55%이다. 따라서 EBB 말하기 평가척도는 전체 수험자를 상, 하 균등하게 이분한다고 볼 수 있다.

표 2
교사 A, B, C의 점수별 합산 빈도

점수*	빈도	점수별 백분율(%)	누가 백분율(%)
3	2	10	10
6	1	5	15
8	2	10	25
9	3	15	40
11	3	15	55
13	2	10	65
14	2	10	75
15	4	20	95
16	1	5	100
합계	20	100	

*교사 A, B, C의 합산점수

두번째로 점수별 세부판정을 알기 위해 EBB 척도개발에 직접 참여한 교사

A, B의 평가내용을 각각 살펴보았다. 표 3은 교사A가 EBB 평가척도를 통해 등급분류한 학생들의 점수이다. 표 3에서 최상위 등급점수인 6점대만 5%로 낮고, 나머지 학생들의 점수는 비교적 균등하게 변별되었다. 바닥효과(floor effect)는 나타나지 않았지만, 5점대에서 30%의 수험자가 몰려 천정효과(ceiling effect)가 약간 나타났다.

표 3
교사A의 평가내용

점수	응시학생*	빈도(비율)
1	정 준, 유정재	2(10%)
2	고이봉, 김정호, 서태호	3(15%)
3	남성진, 봉여대, 손정렬, 차종오	4(20%)
4	김경중, 박태민, 신지호, 한명운	4(20%)
5	김동준, 김주호, 박상빈, 이문수, 이해우, 최성민	6(30%)
6	김기태	1(5%)

*학생들의 이름은 모두 가명임.

표 4에서는 공동 척도개발자 교사B의 평가내용이 나와있다. 교사B의 경우 각 점수대별로 10-20%를 유지하며 균등하게 판정하여 바닥효과와 천정효과가 거의 나타나지 않았다.

표 4
교사B의 평가내용

점수	응시학생	빈도(비율)
1	유정재, 서태호, 정 준	3(15%)
2	고이봉, 김정호, 봉여대, 최성민	4(20%)
3	손정렬, 한명운,	2(10%)
4	김기태, 김주호, 박상빈, 이문수	4(20%)
5	김동준, 남성진, 박태민, 이해우	4(20%)
6	김경중, 신지호, 차종오	3(15%)

끝으로 척도개발 교사 A, B의 수험자 등급결정 내용을 서로 비교분석했다. 교사 A, B가 EBB척도의 하위등급인 1, 2, 3점 판정에는 최대 1점밖에 차이가 나지 않았다. 반면, 상위등급인 4, 5, 6점 판정에는 2-3점까지 견해차이가 나타났다. 교사A가 김기태 학생에게 최고점 6점을 준 반면, 교사B는 김기태 학생에게 4점을 주었다. 교사A가 차종오 학생에게 3점을 준 데 반해 교사B는 6점을 부여했다. 이에 대해 수험자들의 실제 수업담당이었던 교사A는 다음과 같이 그 이유를 설명했다.

> “김경중 학생의 경우 전반적으로 말하기는 우수했으나 수업 중 배운 말하기 과업인 이야기 형식이 아닌 에세이 형식이었습니다. 그래서 육하원

> 칙이 지켜지지 않은 이야기로 판정하여 4점을 주었습니다. 신지호 학생은 봉사활동에 대한 내용을 육하원칙에 의거 자세히 묘사하지 않아 4점을 주었습니다. 차종오 학생의 말하기는 얼핏 들으면 매우 유창해 보였으나 녹음내용을 다시 들어본 결과 처음부터 끝까지 문장과 문장 사이의 휴지(pause) 없이 같은 속도로 내용을 암기하는 것으로 보여 거짓 유창성으로 판단했습니다. 그리고 무슨 내용인지도 이해하기도 어려워 3점을 부여했습니다."(교사A)

연구대상 학생들을 직접 지도하진 않았지만 채점척도를 공동 개발한 교사B는 최고등급 판정에 대해 다음과 같이 말했다.

> "김기태 학생은 내용전달은 좋았지만 억양이 부자연스럽고 한국식 발음이 강해서 그런지 유창해 보이지 않았습니다. 반면 김경중, 신지호, 차종오 학생은 원어민 발음처럼 자연스러웠습니다."(교사B)

교사A에게 최고점수 6점과 실제 수행평가에서도 거의 만점을 받았지만 교사B에게 4점을 받은 김기태 학생의 경우 부록 2와 같이 영어회화 점수 66.9, 듣기평가 점수 13점의 중상의 영어능력을 갖고 있는 것으로 나타났다. 말하기 능력과 듣기, 읽기능력이 차이를 보이는 데에 대해 김기태 학생 본인은 다음과 같이 설명했다.

> "1학년 때는 80점대였는데 2학년에 올라와서 영어점수가 60점대로 떨어졌어요. 2학년에 올라오니 공부해야 할 단어나 문법이 많은데 공부를 좀 소홀히 했습니다. 말하기, 듣기, 독해 중 말하기에 제일 관심이 있었기 때문에 말하기 수행평가는 학원의 도움을 일체 받지 않고, 제가 집에서 혼자 연습해서 시험 본 겁니다. 하지만 제일 자신있는 것은 독해이고 다음으로 듣기, 말하기예요."(김기태 학생)

3. EBB 평가척도 등급판정의 명확성

표 5에는 EBB 평가척도의 등급기준 질문 내용과 해당되는 언어능력의 단면이 제시되어 있다. 부록 1에는 교사A가 등급구분한 말하기 표본 6개가 나와 있다. 부록 1의 말하기 표본과 표 5의 등급기준질문을 대조분석하여 등급판정의 명확성을 살펴보았다.

먼저 최고등급을 받은 김기태 학생의 말하기를 살펴보면 6등급구분 기준인 '부정확한 발음이 있는가?'를 MP3 녹취파일을 통해 분석했다. 분석 결과 김기태 학생의 발음은 'cleaned'라는 단어가 약간 불분명할 뿐 전체적으로 부정확한 발음이 거의 없었다. 그리고 이야기 육하원칙, 이야기 전달력, 과거시제 일관성,

발화중단 여부에 문제가 없었다. 이에 반해 5등급을 받은 이문수 학생은 발화가 한번 중단되고, 휴지(pause)가 많아 전체적으로 발화의 흐름이 자주 끊겨 부정확한 발음으로 판정했다. 그러나 이문수 학생의 이야기 발화에는 'last winter vacation'은 언제, 'Seoul station'은 어디서, 'I'는 누가, 'volunteered'는 무엇을, 'gave out food for homeless people'는 어떻게, 'meaningful works'는 왜로 이야기의 육하원칙이 뚜렷하게 나타났다.

표 5
등급기준 질문

등급		해당영역
6	부정확한 발음이 있는가?	발음
5-4	이야기의 육하원칙이 모두 나타나는가?	이야기 구성력
전체 이분	이야기를 쉽게 이해할 수 있는가?	청자의 이해 가능성
3-2	과거시제를 일관되게 사용하는가?	문법
1	발화가 중단돼 이야기가 마무리가 안 되는가?	담화능력

한편 4등급을 받은 신지호 학생은 'last summer vacation'은 언제, 'handicapped people camp'는 어디서, 'I'는 누가, 'participated'는 무엇을, 'valuable'은 왜로 대체적으로 육하원칙을 표현했다. 그렇지만 봉사활동을 어떻게 구체적으로 했는지에 대한 설명이 부족해 보였다.

3등급을 받은 손정렬 학생의 경우 'saw, was, saw, gave, was, thought' 등 과거시제를 일관되게 사용했다. 2등급을 받은 고이봉 학생은 'want, went to, helped, wish, doesn`t' 처럼 과거동사와 현재동사를 혼용하여, 손정렬 학생 보다 과거시제의 일관성이 떨어졌다. 끝으로 유정재 학생은 발화가 계속 이어지지 못하고, 배경, 사건, 결말의 이야기 구성문법이 나타나지 않아 최저등급을 받았다. 따라서 EBB 평가척도와 6개의 발화표본을 대조분석한 결과, EBB 척도가 대체로 명확한 등급판정을 내리는 것으로 보였다.

한편, EBB 평가척도에 나타난 의사소통능력은 문법, 발음, 청자의 이해가능성, 이야기 육하원칙, 담화능력의 5가지였다. ETS(2001)의 TSE시험 등급 기술표에 의하면 문법, 발음, 청자의 이해가능성은 언어능력에 속하고 이야기 육하원칙, 이야기 구성력은 담화능력에 속한다. 평가척도의 각 기준질문의 분류는 의사소통 모형에 조금 다르게 분류될 수 있다.

그러나 본 연구의 말하기 시험에서 사회언어적 능력이나 기능적 능력 같은 화용적 능력을 평가할 만한 발화특성이 두드러지게 나타나지 않아 평가척도에 반영되지 못했다. 독백이라는 말하기 과업 특성과 초급 수준의 수험자 말하기 능력 변인으로 인해 화용적 능력이 나타나지 않은 것으로 보인다. 말하기 과업이 독백이 아닌 인터뷰, 대화형식이거나 수험자의 언어능력이 높을 경우 EBB 척도에 화용적 능력이 반영될 가능성이 높다고 본다.

4. EBB 평가척도의 채점자간 신뢰도

표 6에 의하면 교사A의 채점평균은 3.60, 표준편차는 1.47이고 교사B는 3.55, 표준편차는 1.73, 교사C는 3.65, 표준편차는 1.63이다. 6점 EBB 평가척도를 사용한 결과 세 교사의 채점분포가 유사했으며, 산포도 확인결과 이상치(outlier)로 판단할 만큼 선형성을 저해하는 점수는 보이지 않았다.

표 6
채점결과의 기술통계

n=20	교사A	교사B	교사C
평균	3.60	3.55	3.65
(표준편차)	(1.47)	(1.73)	(1.63)

표 7에는 ICC 채점자간 신뢰도 계수가 95% 신뢰구간에서 제시되어 있다. 개별채점자인 경우 채점자간 신뢰도가 .54, 평균 채점인 경우 채점자간 신뢰도가 .78로 p<.01수준에서 모두 유의미하게 나왔다. Cronbach alpha 역시 .78로 유의미한 채점의 내적 일관성이 있는 것으로 나타났다. 3명의 채점자간 신뢰도가 다소 낮게 나온 이유는 척도개발에 참여하지 않은 교사C의 참여, 20명의 소표본, 협력부족으로 인한 채점자 표준화 미비 때문으로 추정된다. 따라서 외부교사 참여, 소표본, 평가자 표준화 미비 상황에서 나온 ICC 채점 신뢰도를 고려할 때, EBB 말하기 평가척도는 안정된 척도로서 그 가능성을 보여 준다.

표 7
ICC 채점자간 신뢰도

채점방식	ICC	95% 신뢰구간		참값이 0인 F값	집단간	집단내
		하위	상위		df1	df2
개별 채점	.54**	.28	.76	4.56	19	38
복수 채점	.78**	.53	.91	4.56	19	38

**p < .01, Cronbach alpha = .78

V. 결론 및 제언

본 연구는 Upshur와 Turner(1995)의 EBB 평가척도 모형을 소개하고, 실제 현장적용 사례를 제시하고자 했다. EBB 평가척도를 중등학교에 적용해 본 결과, 수업내용과 평가척도의 연계성, 이분적 등급판정의 명확성, 채점 신뢰성 측면에서 적용가능성이 높아 보였다.

첫째, EBB 평가척도에 교수학습 내용이 비교적 잘 나타나 Upshur와 Turner (1995)의 선행연구 결과를 확인했다. EBB 척도의 총 다섯 등급결정 질문 중 세

개는 교수학습 내용과 직접적으로 관련이 있었으나, 나머지 두 개의 등급구분 질문은 다소 거리가 있었다. 특히 교사B의 인터뷰에서 나타났듯이 EBB 평가척도의 등급결정 질문을 정확한 발음이라는 기존 관행대로 만들 경우, 원어민 중심의 연역적 평가관행을 바꾸기 힘들다는 사실을 발견했다. EBB 평가척도가 원어민 중심의 블랙홀에 빠지기 않기 위해서는 수업 준비단계부터 평가와 연동할 수 있는 교수학습 내용을 명확하게 설정하고 그에 입각하여 말하기 과업을 설계해야 한다(임창근, 1998).

둘째, EBB 평가척도는 이분적 등급판정으로 수험자를 비교적 명확하게 양분했다. 명확한 이분적 등급결정으로 EBB 척도는 비원어민 교사가 감당할 수 있는 실용적 척도로 보인다. 그리고 EBB 척도가 등급기준에 대해 채점자들이 서로 다른 해석을 내릴 가능성을 줄인다는 Upshur와 Turner(1995)의 연구결과를 재확인했다. 그러나 교사 A, B의 최상위 등급판정에서 보듯이 등급결정 기준에 대한 채점자 견해의 표준화가 충분히 이루어지지 않을 경우 채점자들이 서로 다른 해석을 내릴 소지는 여전히 남아있다.

따라서 EBB 척도의 등급결정 과정과 어떤 등급결정 질문을 만드느냐가 EBB 평가척도의 타당성과 신뢰성을 결정짓는다. 척도 공동개발을 통해 채점자들이 공동체를 이루고, 그 채점 공동체에서 교사들이 서로 충분히 논의하고 합의해야 채점자 표준화가 이루어질 수 있다(이영식, 2000). 채점자 표준화 뿐만 아니라 대표 수험자 말하기 표본 결정, 등급결정 질문의 위계성 결정, 이분 알고리즘의 YES/NO 화살표 방향 등 EBB 척도개발은 교사들이 서로 대화하고 협력해야 할 부분이 많다. 이런 지속적인 타당화 과정(Messick, 1989)을 거쳐 만들어진 EBB 척도가 비로소 교육적 측면을 확보하고, 비원어민 교사가 쓸 수 있는 타당한 평가척도가 될 수 있다.

셋째, 본 EBB 척도는 ICC 신뢰도 계수가 단일채점자 .54, 복수채점자 .78로 Upshur와 Turner(1995)의 의사소통능력 척도의 신뢰도 .81과 문법적 정확성 척도 .87보다 다소 낮지만 근접하게 나타나 척도의 신뢰성을 재확인했다. 특히 척도개발에 전혀 관여하지 않은 교사C까지 포함한 채점자간 신뢰도가 .78이란 것은 매우 의미있는 사실이다. 더 나아가 안정된 채점 신뢰성을 바탕으로 동일한 대상, 동일한 수업내용, 동일한 말하기 과업에 대해 개발된 EBB 척도의 재사용도 탐색할 필요가 있다. 정해진 교과수업을 해마다 반복하는 중등학교의 특성상 매번 EBB 척도를 개발하는 것보다, 전년도에 만든 평가척도를 수정보완하여 재사용하는 것이 척도의 실용성을 더 높일 수 있다.

한편, 말하기 시험을 실시한 후, 교사들이 학교현장에서 만든 '평가척도가 주관적이다' 라는 비판이 있을 수 있다. 이에 대해 신동일(2002)은 "(평가의) 타당함의 근거는 바로 연구 공동체에서 개입된 서로의 주관적 또는 전문적 인식의 개입이었다 " (p. 489)라고 상호주관성의 타당성을 역설한다. North와 Schneider(1998) 또한, 평가척도의 위계적 난이도를 평가자의 의견이 아닌 데이터에 근거해 만들었다면 평가척도의 객관성을 확보한 것이라 주장한다. 따라서

학생들을 장시간 관찰한 교사들의 전문적 판단과 학생들의 말하기 표본이 결합된 EBB 평가척도는 객관성을 확보한 것으로 보인다.

반면, EBB 평가척도를 적용해 본 결과 다음의 3가지 제한점이 발견되었다. 첫째, EBB 평가척도는 말하기 과업에 많은 영향을 받는다(오준일, 2006; Chalhoub-Deville, 1995). 이런 과업 의존성과 모집단의 제한성 때문에 EBB 평가척도는 일반화에 제한을 받는다. 둘째, EBB 척도는 평가 후 척도개발로 인한 지연채점으로 많은 채점시간과 노력이 필요하다. 셋째, 동료교사와의 협력 평가를 지원하는 행정 시스템이 필요하다. 영어교사들이 영어 교과부서에서 함께 근무하지 않고 행정부서에서 각자 근무하는 중등학교의 특성상, 협력평가가 자칫 EBB 척도 현장적용의 걸림돌이 될 수 있기 때문이다.

지금까지 본 연구는 중등학교 말하기 평가에서 연역적 평가척도의 문제점을 살펴보고, 귀납적 평가방식인 EBB 평가척도의 현장적용 사례를 통해 타당성과 신뢰성을 검증해 보았다. Fulcher(2007)는 “현재까지 개발된 모든 평가척도와 평가 시스템에 많은 문제가 있다는 것은 타당한 평가척도 개발이 그 만큼 더디고 힘든 길이라는 사실을 보여준다”(p. 16)라고 척도 개발의 어려움을 토로했다. EBB 평가척도 역시 예외는 아니다. 변별을 목적으로 하는 고부담의 큰 시험이 아닌 학습여부를 판단하는 교실상황의 작은 시험이라면 EBB 척도가 그 타당성을 인정받을 수 있을 것으로 보인다. 끝으로 단위학교의 전체 학생들을 대상으로 EBB 척도 확대적용, EBB 척도 등급기술의 위계성과 연역적 평가척도 등급기술의 위계성 비교, EBB 척도의 타당성에 대한 평가자와 수험자의 인식연구, 쓰기평가 EBB 척도개발 등의 후속 연구를 통해 중등학교에서 말하기 평가가 더 활성화되기를 기대해 본다.

참고문헌

교육과학기술부. (2006). *교육통계연보*. 월드와이드웹: http://cesi.kedi.re.kr/index.jsp에서 2007년 11월20일에 검색.

김덕기, 안병규, 오윤자, 김영규. (1999). 중등영어 수행능력 절대평가기준 제안: 표현기능을 중심으로. *영어교육*, *54*(2), 175-200.

김해선. (1999). *중학교 영어 말하기 평가 실태 분석과 제안*. 석사학위논문. 부경대학교, 부산.

박기화. (2005). 7차 영어과 교육과정 성취 기준의 적정성 연구: 말하기, 쓰기를 중심으로. *영어교육연구*, *17*(4), 229-250.

신동일. (2001). ACTFL-SOPI체제의 영어 말하기 평가 현장연구: 한국에서의 문제점과 현실적 대안. *영어교육*, *56*(2), 309-331.

신동일. (2002). Rasch 모형을 이용한 고등학교 영어과 말하기 및 쓰기능력 등급 기술표 개발. *영어교육*, *57*(4), 469-499.

신동일, 서인영. (2002). SLA 선행 연구에 근거한 말하기 평가척도 제안: 문법 영역 중심으로. *영어교육, 57*(2), 423-448.

신상근. (2007). 예비교사의 평가 전문성 신장을 위한 동료평가의 유용성에 관한 연구. *영어교육연구, 19*(1), 183-200.

오준일. (2006). 영어 말하기 능력 절대 등급 설정에 관한 연구. *영어교육연구, 18*(1), 141-162.

이병민. (2003). EFL 영어학습 환경에서 학습시간의 의미. *Foreign Languages Education, 10*(2), 107-129.

이영식. (2000). 영어 작문평가에 대한 채점자 훈련의 원리. *영어교육,55*(2), 201-217.

이완기. (2003). *영어 평가 방법론*. 서울: 문진미디어.

이인제, 이범홍, 박정, 진재관, 김옥남, 서수현, 김신영, 전병만, 박준언, 안병규, 오준일, 유제명, 이소영, 김신혜. (2004). *영어과 교사의 학생 평가 전문성 신장 모형과 기준*. (연구보고 RRE 2004-5-7). 서울: 한국교육과정평가원.

이 호. (2005). 영작문 평가 상황에서 학생들의 평가 과정에 대한 연구. *영어교육연구, 17*(3), 215-234.

이 호. (2007, 12월). *시험문항 지침서를 통한 예비교원 평가훈련*. 한국외국어교육학회 학술대회 발표논문, 서울.

임창근. (1998). 영어교육 평가에 있어서 질적인 평가와 양적인 평가의 조화. *초등영어교육, 4*(1), 167-187.

전병만, 박준언, 유제명, 최희경. (2006). *초·중등 영어교육 현황분석*. (교육부 정책과제 2006-이슈-4). 서울: 교육과학기술부

최인철. (2005). 중등영어교육 수행평가를 위한 음성인식기술기반 모의구술면접모형. *Foreign Languages Education, 12*(4), 235-266.

ACTFL. (1986). *ACTFL proficiency guidelines*. Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Adams, M. L., & Frith, J. R. (1979). *Testing kit: French and Spanish*. Washington DC: Department of State and the Foreign Services Institute.

ALC. (2007). *SST 등급해설*. 월드와이드웹: http://www.alc.co.jp/edusys/sst/level.html에서 2007년 11월17일에 검색.

Atkinson, J. M., & Heritage, J. C. (Eds.). (1984). *Structures in social action: Studies in conversation analysis*. Cambridge: Cambridge University Press.

Bachman, L. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*, 149-164.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal, 70*, 380-390.

Brewer, M. B. (2000). Research design and issues of validity. In H. T. Reis & C. M. Judd

(Eds.), *Handbook of research methods in social and personality psychology* (pp. 3-16). Cambridge: Cambridge University Press.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards & W. Richards (Eds.), *Language and Communication* (pp. 1-12). London: Longman.

Canale, M., & Swain, M. (1980). The theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics, 1*, 1-47.

Carroll, B, J. (1982). Language testing: Is there another way? In J. B. Heaton (Ed.), *Language Testing* (pp.1-10). London: Modern English Publications.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*, 16-33.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Davidson, F., & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

E.T.S. (2001). *TSE and SPEAK score user guide*. Retrieved December 20, 2007, from the World Wide Web: http://ftp.ets.org/pub/toefl/008659.pdf.

Finch, A., & Shin, Dong-Il. (2005). *Integrating teaching and assessment in the EFL classroom*. Seoul: Sahoe Pyeogron Academy.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *English Language Teaching Journal, 41*(4), 287-291.

Fulcher, G. (1988). *Lexis and reality in oral testing*. Washington, DC: ERIC Clearinghouse on Languages and Linguistics. (ERIC Document Reproduction Service No. ED 298 759).

Fulcher, G. (1996). Does the thick description lead to smart tests?: A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Fulcher, G. (2007). Evaluating quality in second language performance tests. *English Language Assessment, 1*, 3-19.

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics, 25*, 205-227.

Hwang, Sung-Sam. (2007). Intrarater reliability of speaking tests in high school: A focus on rater's overfit rating patterns. *TESOL Forum, 1*, 93-105.

Khattri, N., Reeve, A., & Kane, M. (1998). *Principles and practices of performance assessment*. Mahwah, NJ: Erlbaum.

Lantolf, J. P., & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *The Modern Language Journal, 69*, 337-345.

Lazaraton, A. (2002). *Studies in language testing 14: A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

Liskin-Gasparro, J. E. (1996). Circumlocution, communication strategies, and the ACTFL proficiency guidelines: An analysis of student discourse. *Foreign Language Annals, 29*, 317-330.

Lowe, P. (1985). The ILR oral interview: Origins, applications, pitfalls, and implications. *Die Unterrichtspraxis, 16*, 230-244.

Magnan, S. S. (1987). Rater reliability of the ACTFL oral proficiency interview. *The Canadian Modern Language Review, 43*, 267-276.

Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *English Language Teaching Journal, 44*(2), 117-121.

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

North, B. (1994). *Scales of language proficiency: A survey of some existing systems*. Strasbourg: Council of Europe.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-263.

Pienemann, M., Johnson, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition, 10*, 217-234.

Shohamy, E. (1990). Language testing priorities: A different perspective. *Foreign Language Annals, 23*(5), 365-394.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics, 15*, 188-211.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Essex: Pearson Longman.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing, 1*, 202-220.

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *New directions in discourse processing* (pp. 53-120). Norwood, NJ: Ablex.

Upshur, J., & Turner, C. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*, 3-12.

Upshur, J., & Turner, C. (1999). Systemic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 82-111.

Wilds, C. (1975). The oral interview test. In R. L. Jones & B. Spolsky (Eds.), *Testing Language Proficiency* (pp. 29-44). Arlington, VA: Center for Applied Linguistics.

부록 1

교사**A**가 등급 판별한 6명의 말하기 표본 전사자료

영어회화 과목 말하기 수행평가 – M고등학교 - 2007년 6월- 전사: 황성삼

Level 1 - 유정재

1 I'm going to help(.)ah I'm going to Africa to help the pool I will 2 the people(.)who is pool I watched TV

Level 2 – 고이봉

1 I want I went to a(.)luck) home for home for orphans I helped child
2 childrens looked like happy my feel was good for (one) I serves
3 I served(.)was good I served was good hard to volunteering
4 my heart was (broke) I wish that children's smile doesn't lose and
5 I will serve serves to volunteering can you join us?

Level 3 - 손정렬

1 when I saw(.)was surfing internet I saw a picture of a she gave
2 her muffler to homeless people in in a subway station seeing the
3 picture I(.)was very impressed she gave her muffler
4 without without any benefit for her I think(.2)I thought that was
5 example of volunteer work volunteer work mean that that he or she
6 do that for other she was really doing volunteer work if
7 if we consider other we we can do volunteer work I think I think
8 you we should think other and we should volunteer work we if we
9 everyone every one does volunteer work the world will be better

Level 4 - 신지호

1 CA*:I participated in a handicapped people camp(.)at the last
2 summer vacation it was a new experience for me I ah never
3 participated because I had in a camp with the handicapped so
4 first I swore oh I swore to do volunteering work eager but it
5 was not easy most of the handicapped(.)were not communicated
6 with me I was felt an emotion of pity after the volunteering
7 was finished I thought this work was very valuable
8 IN:so I will try?
9 so I will try to volunteer
10 CA:to help others whenever I have leisure times

Level 5 - 이문수

1-> CA: last winter vacation I will I volunteered at the I helped homeless
2 people at Seoul station I gave gave out food for
3 homeless people I ah! felt 처음부터 다시 하면 안될까요?
4 IN: volunteering work.
5 CA: ah ye volunteering volunteering work was so hard(.2)but
6 volunteering work is was meaningful work so I am going
7 to volunteer(.)volunteer(.)at a animal shelter in this
8 summer vacation I like to take care of animals(.)so ah therefore
9 I will do well probably probably this work will be amusing
10 IN: ok. 다음 사람?

Level 6 - 김기태

1 last week I did volunteer work(.)with my friend we went to the fire 2 station there were many fire engines(.)I washed a car and
3->my friend cleaned up office it was too difficult for me
4 It took 3 hours when we finished it we felt pleased
5 a fire fighter gave us a cup of coffee the coffee was so delicious 6 when I came back home(.)I was so tired

*CA:candidate, IN:interviewer 그 밖의 전사는 Atkinson과 Heritage(1984)를 따름.

부록 2

대표 수험자 8명의 1학기 영어회화 점수 및 듣기평가 점수

학 생	영어회화 점수 (100점 만점)	듣기평가 점수 (17점 만점)
신지호	92.2	15
이문수	76.3	16
김기태	66.9	13
김주호	58.2	12
김정호	43.5	11
정 준	35.3	10
유정재	31.8	9
봉여대	26.5	11

교육단계(Applicable levels): 중등교육, 초등교육
주제어(Key words): 말하기 평가, EBB 교실평가, 평가척도

황성삼
중앙대학교 영어교육과
156-756 서울특별시 동작구 흑석동 221
Email: ecloguehwang@hotmail.com

Received in February, 2008
Reviewed in March, 2008
Revised version received in May, 2008