10.15738/kjell.17.4.201712.787

한국 영어 학습자의 유창성과 발음에 대한 비원어민 채점자의 이해 및 판단 기준 분석*

송민영
(고려사이버대학교)

Song, Min-Young. 2017. Nonnative raters' perceptions and judgments of Korean English learners' fluency and pronunciation level. *Korean Journal of English Language and Linguistics* 17-4, 787-815. This study aims to investigate how nonnative English raters understand the concept of 'fluency' and 'pronunciation', and what criteria they apply in evaluating Korean English learners' speaking performance. Most of the participants in this study understood the main concepts of fluency defined in NEAT and used them in oral reports. However, rather than judging the absolute levels of fluency applying those concepts, they determined fluency scores considering other factors such as task completion, amount of utterance or the level of the test concerned. Especially, they tended to judge fluency levels in relation to task completion levels. As for pronunciation, most of the participants did not consider the main concepts of pronunciation defined in NEAT. Instead, they applied 'intelligibility' as the only crucial criterion in determining pronunciation scores. In particular, some participants applied this criterion too generously to discriminate pronunciation levels appropriately. Also, many participants tended to adjust pronunciation scores based on task completion scores. In conclusion, although the participants of this study were relatively reliable and experienced scorers, most of them did not apply appropriate scoring criteria in evaluating Korean students' English fluency and pronunciation levels.

Keywords: fluency, pronunciation, nonnative raters, rater verbal reports, analytic scoring, speaking assessment

1. 연구의 필요성 및 목적

우리나라 영어교육의 목표가 네 언어 기능을 고르게 성장시키는

* 본 논문은 2013년 한국교육과정평가원이 발간한 '국가영어능력평가시험 채점 신뢰도 및 타당도 향상 방안' 보고서의 일부 내용을 재구성 및 보완한 것임.

데 있음에도 불구하고 오랜 기간 중·고등학교 교육 현장에서는 말하기와 쓰기를 제대로 지도하고 평가하는 일을 소홀히 한다는 비판을 받아왔다. 지난 이명박 정부에서는 NEAT(National English Ability Test)라는 영어의 네 기능 영역을 모두 평가하는 시험을 개발하여 고등학생에게 적용함으로써 수업에서 표현기능(productive skills)도 제대로 지도하도록 유도하는 평가의 긍정적 환류효과(beneficial washback)를 기대하였으나, 박근혜 정부에서는 결국 이 시험을 폐지하기로 결정하였고 그 이후로 이제까지 공교육에서의 영어 말하기·쓰기 평가에 관한 논의가 그리 활발하지 않은 것이 사실이다.

NEAT의 전면 중단으로 인해 그 도입을 준비하면서 축적되었던 정보 역시 제대로 활용되지 못하고 있지만 네 언어 기능을 고르게 발전시킨다는 영어교육의 목표가 여전하고 전문가 및 현장 교사들의 노력이 계속되는 한 그러한 자료는 의사소통능력 중심 영어교육 시행 및 발전에 중요한 자료로 활용될 수 있을 것이다. 특히, 현직 중·고등학교 영어교사들을 NEAT 채점자로 활용하기 위해 시행되었던 연수는 우리나라 공교육 내에서 이뤄졌던 최초의 말하기·쓰기 채점 연수로서 이를 통해 얻은 자료는 NEAT에만 적용되는 것이 아니라 영어교사들의 말하기와 쓰기 지도 및 평가 전문성 신장을 위한 연수에서 활용될 수 있을 것이다.

본 논문은 NEAT 채점자 연수에 참여하였던 영어교사들 중 우수한 채점자들의 채점 행동의 특성을 파악하여 이를 채점자 훈련에 적용하려는 목적으로 시행되었던 연구의 일부로서, 우리나라 영어 학습자를 대상으로 분석적 채점을 적용하는 말하기 시험에서 영어 비원어민으로서 채점에 참여하는 교사들이 각 영역 점수를 어떻게 결정하는 지 파악하는 데 초점을 둔다. NEAT는 공교육 내에서 실행되는 평가로서 학생들에게 말하기 세부능력별 피드백을 제공하기 위해 분석적 채점 방식을 채택하였는데 음성답안을 들으면서 각 영역을 독립적으로 채점하는 것은 충분한 훈련이 필요한 일이다.

Song과 Lee(2015)의 연구에서는 NEAT 채점자 연수에 참여했던 교사들 중에 기준 점수와의 일치도가 우수한 말하기 채점자들과 그렇지 못한 채점자들의 전반적인 채점 행동을 비교하여 채점의 일관성을 높이기 위해 필요한 전략을 파악하고자 하였다. 본 연구에

서는 해당 연구에 참여했던 우수한 채점자들이 분석적 채점에서 특정 영역의 채점 기준을 어떻게 이해하고 적용하는 지를 구체적으로 파악하고자 하였다. 이전 연구가 채점의 일관성을 높이기 위해 사용하는 전략을 파악하는 데 초점을 두었다면, 본 연구는 동일한 채점자들이 각 채점 영역을 어떻게 이해하고 있으며 점수대별로 어떤 판단 기준을 적용하는 지, 즉 채점의 타당도(validity)를 확보하고 있는 지를 파악하는 데 초점을 두었다. 특히, NEAT 말하기의 다섯 가지 채점 영역(과제완성, 유창성, 발음, 구성, 언어사용) 중 '유창성(fluency)'과 '발음(pronunciation)' 영역에서 채점자들이 점수대별로 어떤 기준을 적용해 점수를 정하는 지를 파악하고자 하였다. 유창성과 발음 두 영역을 택한 것은 우선 연구 참여자들 대부분이 가장 채점이 어려운 영역으로 5점 척도를 적용하는 유창성을 꼽았고 점수대별 채점 기준이 모호하다는 점을 지적했기 때문이다. 이와 비교할 때 3점 척도를 적용하는 발음은 비교적 일관성 있게 채점하는 것이 상대적으로 용이하지만 참여자들의 채점 기준 발언을 분석하는 과정에서 채점의 일관성과는 별개로 발음과 유창성의 개념을 혼동하는 등 채점의 타당성이 의심스러운 사례가 다수 나타나 유창성과 더불어 심층적인 분석이 필요하다고 판단했기 때문이다.

본 연구의 연구 질문을 구체적으로 기술하면 다음과 같다.

1. 영어 말하기 시험에서 비원어민 채점자들은 유창성을 어떻게 이해하고 점수를 정하는 데 어떤 기준을 적용하는가?
2. 영어 말하기 시험에서 비원어민 채점자들은 발음을 어떻게 이해하고 점수를 정하는 데 어떤 기준을 적용하는가?

2. 선행 연구

2.1 L2 말하기 채점의 타당성

외국어 말하기 평가에서 채점자가 어떤 기준을 적용해 점수를 판단하는 지, 즉 점수 판정의 타당성을 확보하고 있는 지는 채점의 일관성과 더불어 논란의 핵심이 된다. 이와 관련해 질적

(qualitative) 분석 방법을 적용해 채점자들의 점수 판단 과정에 대한 분석을 시도한 많은 연구들이 있었다(예, Brown 2005, Kim 2010, Orr 2002). 채점 기준에 대한 채점자의 해석을 분석한 연구에 따르면, 채점자들은 채점 루브릭의 서로 다른 특정 요소에 과도하게 집중하기도 하고(예, Chalhoub-Deville 1995, Merion 1998, Orr 2002), 정해진 채점 루브릭과 관계없는(construct-irrelevant) 기준을 적용하기도 하고(예, Kim 2010, Meiron 1988, Papajohn 2002), 완전히 다른 이유로 같은 점수를 부여하기도 하는 것으로 보인다(예, Meiron and Schick 2000, Orr 2002). 즉, 채점자들은 동일한 채점 연수를 거친 후에도 여전히 주어진 채점 루브릭에 대해 다르게 이해하고 점수 판단에 적용하는 기준이 다르기도 하다.

이런 채점자간 불일치의 주요 원인으로 채점자들의 배경 요인이 다르다는 점에 주목하여 구체적으로 어떤 차이가 있는 지를 비교하는 연구들도 다수이다. 예를 들어, 채점의 숙련도에서 차이가 있는 집단의 채점 행동을 비교하거나(예, Kim 2010, Song and Lee 2015), 원어민 채점자와 비원어민 채점자를 비교하거나(예, Kim 2009, Zhang and Elder 2011), 해당 언어를 지도한 경험이 있는 교사와 교사가 아닌 채점자들을 비교하기도 하였다(예, Brown 1995, Meiron and Schick 2000). 이와 비교할 때 본 연구에 참여하는 채점자들은 모두 우리나라 중·고등학교 학생들을 가르치는 중견 영어교사로서 비교적 영어 학습의 경험이 비슷한 EFL 영어 학습자들이다. 뿐만 아니라 말하기 채점자로서의 경험과 숙련도 역시 크게 다르지 않다. 따라서 본 연구는 이전 연구들과는 달리 비슷한 배경을 가지고 있는 영어 말하기 채점자들이 채점 기준을 이해하고 적용하는데 어떤 공통점과 차이점을 보이는 지 비교하는 연구로서 의미가 있을 것이다.

2.2 L2 유창성 및 발음에 대한 평가

L2 말하기 평가에서 유창성에 관한 연구의 주된 관심 중 하나는 유창한 발화를 예측할 수 있는 '시간 변인(temporal variables)'이 무엇인가에 관한 것이다(예, Bosker et al. 2013, Derwing, Rossiter, Munro and Thomson 2004, Ginther, Dimova and Yang

2010). 유창성에 대한 인식은 언어별로 차이가 있기 때문에(예, Campione and Véronis 2002, Raupach 1987), 영어를 대상으로 한 연구에만 국한시켜 보면 Kormos와 Denes(2004)는 'speech rate', 'mean length of runs', 'phonation−time ratio', '# of stressed words/min.', 이렇게 네 개의 변인이 유창성 점수를 가장 잘 예측했고 'pause의 빈도'는 의미 있는 예측 변수가 되지 못했다. 반면에, Derwing 외 3인(2004)에서는 유창성 점수와 'speech rate', 'pause'와의 상관관계가 높았고 시험 과제 중 'narrative'는 'monologue'와 'dialogue'에 비해 유창성 점수가 낮았으며 '이해가능정도(comprehensibility)'와 유창성 점수의 상관관계가 높았다. 마찬가지로 Rossiter(2009)에서도 'unfilled pause의 평균 빈도', 'speech rate' 두 변인이 유창성 점수와 상관관계가 높았고 'repetition', 'filler' 등은 부정적인 관계를 보였다. Ginther외 2인(2010)에서는 'speech rate', 'speech time ratio', 'mean length of run', 'the number & length of silent pause'라는 변인들이 유창성 점수와 상관관계가 높았다. 종합하면, 전반적으로 발화의 속도(speech rate), 휴지 없이 말하는 발화의 길이, 휴지(pause)의 빈도 또는 길이 등이 유창성 점수와 관련된 주요 시간 변인이고 그 외에 말하기 과제의 종류나 이해가능정도라는 발음 관련 변인도 유창성 점수와 관련이 있는 것으로 보인다. 다만, 이러한 변인들이 영어 유창성의 주요 개념을 이루고 점수를 예측하는 주요 요인이지만 실제 채점자가 유창성 점수를 판단하는 데 어떤 기준을 적용하는지 채점자의 구두보고를 통해 파악한 사례는 거의 없는 것으로 보인다.

외국어 발음 평가에 미치는 채점자 요인에 관한 선행연구에서는 주로 원어민 채점자와 비원어민 채점자를 비교하였다(예, Brown 1995, Lee, Lim and Kim 2014, Yun 2009a). Brown(1995)은 일본어 말하기 시험에서 원어민 채점자와 비원어민 채점자를 비교했는데 전반적인 평가 점수에서는 차이가 없었으나 발음 영역에서 비원어민 채점자가 더 엄격하게 채점한 것으로 나타났다. 반면에, Kim(2009)에서는 두 집단이 채점 결과에서는 통계적으로 유의미한 차이가 없었으나 영어 원어민 채점자들이 발음을 포함해 각 채점 영역 평가 기준에 대해 비원어민 채점자들보다 훨씬 자세하게

기술했다. 최근 연구인 Lee 외 2인(2014)에서도 두 집단의 채점 결과는 유의미한 차이가 없었으나, 구술보고를 통한 비교에서 영어 원어민 채점자들은 자음과 모음의 발음, 강세에 초점을 두는 반면 한국인 채점자들은 자음을 제외하고 다른 요소에는 주목하지 않았다. Yun(2009a, 2009b)에서는 리듬, 억양, 이해도 등과 같은 발음 구성 요소에 대한 판단에서 두 집단을 비교했는데 한국인 채점자들은 /l/, /r/과 같은 자음이나 우리말에서 구분이 어려운 모음의 발음을 지각하는데 취약했다. 결국, 비원어민 영어 화자가 채점자로서 영어 발음을 채점할 경우 영어 원어민 채점자와는 다른 기준을 적용하는 것으로 보인다. 본 연구에서는 Lee 외 2인(2014)의 연구에서와 마찬가지로 구술보고를 통해 한국인 영어교사 채점자들이 한국인 학생들의 영어 발음 판단에 어떤 기준을 적용하는 지 파악하고자 한다.

3. 연구 방법 및 절차

3.1 연구 참여자

본 연구에 참여한 말하기 채점자들은 한국교육과정평가원에서 실시한 NEAT 말하기 채점자 연수를 이수하고 2012년과 2013년 사이에 실시된 NEAT 베타 시험에 응시한 전국 고등학교 3학년 학생들의 말하기 답안을 채점한 경력이 있는 현직 영어교사들이다. 채점자 인력풀에서 평가원이 제공한 기준 점수와의 일치도, 즉 채점 신뢰도를 기준으로 연구 참여 의사를 밝힌 상위 수준 채점자(0.75 이상) 6명을 대상으로 심층면담을 실시하였다. 표 1에 제시한 연구 참여자들의 세부적인 배경 정보에서 알 수 있듯이, 상위 채점자들이 전반적으로 NEAT 채점 횟수가 많고 교사 경력이 길었으며 모두 1급 정교사의 신분이었다.

표 1. 연구 참여자 배경 정보

ID	연령	채점 신뢰도	NEAT 채점 횟수	영어교사 경력	교사 신분
SH-1	40대	.8482	3	18년	고등학교 1급 정교사
SH-2	40대	.9078	4	14년	고등학교 1급 정교사
SH-3	30대	.8387	2	23년	중학교 1급 정교사
SH-4	50대	.7524	4	29년	고등학교 1급 정교사
SH-5	30대	.7740	3	14년	고등학교 1급 정교사
SH-6	30대	.7542	3	11년	고등학교 1급 정교사

출처: Song and Lee 2015, p. 1085에서 발췌 후 편집

3.2 연구 방법 및 절차

3.2.1 채점자 심층면담

연구 참여 채점자와의 심층면담은 NEAT 베타 시험 채점 합숙 중 이루어졌고 다음과 같은 절차로 진행되었다. 우선 심층면담 참여자 한 명씩 채점용 노트북을 지참하고 연구자가 있는 방으로 들어와서 자신의 일반적인 채점 방식 또는 전략에 대해 면담을 진행했다. NEAT 말하기 각 문항은 '과제완성(1～5점)', '유창성(1～5점)', '발음(1～3점)', '구성력(1～3)', '언어사용(1～5)', 이렇게 다섯 가지 채점 영역에 대해 분석적 채점(analytic scoring)을 하도록 되어있는데 채점 시 답안별로 다섯 가지 채점 영역에 관해 모두 채점을 하는 지 아니면 자신에게 할당된 모든 답안에 대해 하나의 채점 영역을 먼저 채점하고 다른 채점 영역으로 넘어가는지, 그 외에도 채점을 위한 자신만의 특별한 방식이 있는 지, 채점이 가장 어려운 영역이 있는 지 등을 보고해 달라고 요청했다.

그 다음은 연구자와의 면담을 위해 방으로 들어오기 이전에 미리 채점을 마친 대략 5~6 개 정도의 답안에 대해 채점 영역별로 왜 해당 점수를 부여했는지 구체적으로 설명하도록 유도함으로써 '사후구술보고(retrospective verbal protocol)'를 진행하였다. 예를 들어, 연구자는 "왜 이 답안의 유창성 점수는 4점인가요?", "왜 이 답안의 발음 점수는 2점이고 이전 답안은 1점인지 비교해주실 수 있나요?" 등과 같이 질문을 하고 채점자가 가급적 그 이유를 구체적

으로 기술하도록 유도하였다. 이 때 채점자는 설명에 앞서 또는 설명 중간에 해당 답안을 원하는 만큼 다시 들어보고 점수 판단의 근거를 설명할 수 있도록 하였다. 이때 연구자는 자신의 질문이 채점자의 설명에 영향을 미치지 않도록 주의하였다.

이와 같이 이미 채점을 끝낸 답안에 대한 점수 부여 근거를 묻고 답하는 절차가 끝나면 연구자가 지켜보는 상황에서 약 다섯 개 정도의 새 답안을 채점하고 즉석에서 해당 답안의 채점 영역별 점수 산출 근거를 설명하도록 하였다. 각 채점자와의 심층면담은 대략 40~50분 정도가 소요되었고 모든 연구자의 질문과 채점자의 구술 보고는 사후 분석을 위해 녹음 및 저장되었다. 비록 채점자마다 비슷한 숫자의 답안에 대해 영역별 점수 판단 근거를 묻긴 했지만 채점자에 따라 영역별 판단 근거 설명이 모호한 경우가 있어 개인별로 유효한 보고 내용의 양이 동일하지 않았고 아래 연구 결과에서 보듯이 채점자별 분석 답안의 숫자가 상이하였다.

3.2.2 면담 자료 분석

채점자와의 심층면담 자료는 국내 한 사범대학 영어교육과 대학원 재학생 다섯 명에 의해 전사되었다. 전사된 자료는 각 채점자가 답안별로 말한 내용을 정리하였고 이 정리된 자료는 다시 5개 채점 영역(과제완성, 유창성, 발음, 구성, 언어사용)을 키워드로 코딩 작업을 하였고 표 2에 제시된 분석 범주를 바탕으로 각 발화를 분석하였다. 이렇게 함으로써 채점자별 각 채점 영역의 정의 및 점수대별 채점 기준에 대한 이해, 채점 기준 적용 방식 등을 파악하고자 하였다. 본 연구에서는 그 중 '유창성'과 '발음' 영역에서 채점자의 채점 기준 이해와 적용 방식을 파악하는 데 초점을 둔다.

표 2. 전사자료 분석 범주

상위 범주	하위 범주
채점 영역 이해	• 각 채점 영역의 구인(construct)에 관한 이해 • NEAT 채점 영역 정의와 비교 • 문항유형별 채점 영역에 대한 이해
채점 기준 적용	• 영역별 채점 기준 적용 방식 • 채점자별 채점 기준 적용의 특징 • 문항유형별 채점 기준 적용의 차이

4. 연구 결과 및 논의

4.1 채점자들의 유창성 이해 및 점수 판단 기준

연구 참여 채점자들은 유창성 영역에서 5점 척도로 세부적인 점수 결정을 해야 하는 데 점수대별로 적용해야 하는 판단기준을 정하기 어렵기 때문에 주로 '과제완성' 채점 영역에 종속해서 판단하는 경향을 보였다. '과제완성' 영역에 대한 종속은 사실 NEAT 개발진이 채택한 소위 '골든룰' 원칙 때문인데, 한 답안에 대해 '과제완성' 영역의 점수를 정한 다음 이에 따라 '유창성'과 '언어사용' 영역 점수의 상한선을 제한하는 것이다. 즉, 과제완성을 제대로 못하여 발화의 양이 적거나 과제와 관련 없는 발언을 한 경우 아무리 유창하고 언어적 오류가 적더라도 두 영역에서 최고점을 받을 수 없도록 상한선을 제한하는 것이다. 이와 같이 유창성 점수 판단은 과제완성 점수의 영향을 받기 때문에 이후 모든 예시 자료에는 두 영역의 점수를 함께 제시한다.

채점자 SH-1은 표 3에서 보듯이 유창성을 판단할 때 '발화의 양'을 고려하고 '흐름의 연결', '끊어지지 않는 정도'를 고려하고 "너무 느린"과 같은 '발화의 속도' 면에서 판단하는 듯하다. 이 중 '발화의 양'을 제외하고 다른 개념들은 실제 NEAT의 유창성 개념을 구성하는 주요 요소인 'natural rate of speech', 'proper pauses based on thought groups'와 상통한다. 하지만 왜 유창성 점수가 3점인지, 4점인지에 대한 판단 근거는 과제완성 점수에 따른 골든룰 적용 후 그 수준에서 괜찮은지, 약간 부족한지에 대한 판단으로 점

수를 결정하는 것이지 유창성의 개념 내에서 절대적으로 적용하는 기준을 언급한 예는 없었다. 특히, 과제완성이 5점인 경우 유창성 점수에 상한선이 없어 그 질적 판단을 제대로 해야 하는 상황인데 “너무 느린 것 같아서”라고 하며 4점을 주었다. 실제 NEAT에서 유창성 4점은 “answered generally at a natural rate of speech and made proper pauses”라고 하여 ‘너무 느리다’는 판단과는 차이가 있다. 종합하면, SH-1은 유창성의 기본 개념은 제대로 이해하지만 구체적인 점수대별 판단 기준을 내재화하여 적용하지 못하고 과제완성 수준에 따른 상대적인 판단과 발화의 양을 고려하여 최종 점수를 정하는 것으로 보였다.

표 3. SH-1의 유창성 점수 판단 및 보고 내용

과제완성	유창성	보고 내용
3	3	이 정도의 유창성을 가지고 발화량이나 이런 거 했을 때 나름 흐름을 가지려고 노력을 해서 유창성은 3점을 줬을 거 같구요.
3	3	<중략> 발화량이 적은 상태에서... 이정도의.. 어떻게 말해야 되지.. 음 자기가 말한 발화량에서 이 정도 약간 흐름이 좀 연결 된다, 그 정도, 말하려고 하는 걸 그래도 표현을 한다 ,너무 심하게 끊어지지 않고, 그 정도가 3점?
5	4	유창성에선 조금 너무 느린 거 같아서 깎았고...

채점자 SH-2는 본 연구에 참여한 상위 채점자들 중에서 유일하게 0.9가 넘는 채점 신뢰도를 보였고 NEAT 베타 시험에 네 번이나 참여했던 우수 채점자였다. 그리고 표 4의 보고 내용을 분석해 보건데 이 채점자는 채점의 타당성도 상당 수준 확보하고 있는 것으로 보인다. SH-2는 유창성 점수를 결정할 때 과제완성에 따른 상한점수를 고려하되 그 안에서는 절대적인 점수 기준을 적용하려는 경향을 보인다. 주로 pause를 판단 기준으로 삼는데 pause가 길면 “very unnatural” 하여 2점으로 판단하고, pause가 있지만 의사소통을 해칠 정도가 아니라면 “somewhat unnatural” 하다고 하여 3점으로 판단한다. 또한 같은 4점이라도 하나는 “pause가 많은 편”이지만 그것이 “상대방의 attention을 계속 유지할 만큼”이라고 하였는데 이는 NEAT에서 유창성 4점을 정의하는 “generally proper pauses based on thoughts groups”와 부합하는 판단이라고 할 수

있다. 하지만 다른 유창성 4점은 그 자체로서는 “상당히 괜찮은”데도 불구하고 과제완성 점수가 정한 상한선을 넘을 수 없기 때문에 4점으로 결정했다. 이 예에서 보듯이 이 채점자는 골든룰을 지키면서도 그 안에서 유창성의 절대적인 수준을 판단하고 있는 것이다. 그 밖에 이 채점자는 “redundancy”를 고려하였는데, 여기서 redundancy가 많다는 것은 같거나 유사한 단어나 표현을 자꾸 반복하는 경우를 의미할 것인데, 사실 이 개념은 NEAT 유창성의 주요 개념에는 포함되지 않는다. 그럼에도 불구하고 redundancy는 말의 자연스런 흐름을 방해할 수 있어서 NEAT의 유창성 3점 기준인 “somewhat unnatural speed of speech”라는 특징과 통한다고 볼 수 있을 것이다. 이렇게 SH-2는 ‘발화의 속도(rate of speech)’에 관한 직접적인 언급은 없었지만 pause의 길이와 숫자를 근거로 말의 흐름이 자연스러운 정도로 유창성 개념을 이해하고 그 정도에 따라 점수대별 채점 기준을 적용하는 것으로 보인다. 이와 같이 SH-2는 본 연구에 참여한 채점자들 중에 유일하게 유창성의 점수대별 판단 기준을 내재화하여 채점에 적용하고 있는 경우였다.

표 4. SH-2의 유창성 점수 판단 및 보고 내용

과제 완성	유창성	보고 내용
2	2	워낙 pause라든지 이런 것들이 너무 길고 그래서 very unnatural하다라고 생각이 돼서
4	3	일단은 pause가 꽤 있었지만 커뮤니케이션을 너무 지나치게 해칠 정도는 아니고, 그래서 somewhat unnatural정도? 그래서 3점을 줬고요. [유창성은 3점 만점인가요?] 아니요 5점이요. 이제 근데 이거 같은 경우는 이제 과제(완성)가 있기 때문에 그 개수에 영향을 유창성이 받으니깐 이 4점에 충분히 다 퍼펙트하게 맞으면 4점을 주겠지만 그 사이에 pause도 상당히 많았고 그래서 3점을 줬고요.
4	4	상당히 지금 pause가 지금 많은 편인데, 그치만 계속 그 상대방이 attention을 계속 유지할 만큼은 지금 유창성을 유지하고 있다라고 지금 판단이 돼서 general하다라고 봤고.
4	4	유창성도 상당히 괜찮은 아이인데요, 그치만 지금 이 4점을 넘기가 어려운 것 같아요. 그래서 과제완성에 따라서 지금 4점 들어가고.
5	3	4점하고 3점하고 고민을 좀 했었는데, 애 같은 경우는 좀 redundancy가 너무 많다라는 생각이 들었구요.

* [ ]안의 내용은 연구자의 질문 내용
** ( )안의 내용은 문맥 보충을 위해 연구자가 추가한 것임

채점자 SH-3의 경우 실제 답안 채점에 대한 설명에서는 유창성 관련 발언의 양이 많지 않아 점수 판단 기준을 파악하기가 어려웠다. 표 5에서 보듯이, 이 채점자는 일반적인 채점 방식 및 전략에 관한 응답에서 NEAT 말하기 문항 중 6장의 그림을 묘사하여 하나의 스토리를 완성하는 '그림묘사하기' 문항을 예로 들어 설명을 하는 데, 유창성이 과제완성에 종속되는 경향이 있기도 하지만 그림 두 장을 합쳐서 대충 묘사하는 경우 과제완성에서는 감점을 하더라도 유창성에서는 전체적인 흐름을 보고 채점한다고 했다. 이 채점자는 "첫 번째 문장과 두 번째 문장에서는 자연스럽게" 이어지는 경우로 유창성을 '문장들 간의 자연스런 흐름'으로 이해하고, 다른 예에서는 pause로도 정의했다. 하지만 SH-1과 마찬가지로 과제완성의 수준에 따른 상대적인 비교이지 유창성 자체의 절대적인 점수 판단 기준을 적용하는 예는 찾을 수 없었다. 과제완성이 5점인데 유창성을 4점을 준 경우 "pause가 막 들어가는"데도 불구하고 4점

이라고 판단한 것은 과제완성 수준에 따른 상대적인 판단으로 보인다. 종합하면, SH-3은 NEAT의 유창성 개념은 제대로 이해를 하고 있지만 점수대별 절대적인 판단 기준을 내재화하지 않고 과제완성 수준에 따른 상대적인 판단을 하는 것으로 보인다.

표 5. SH-3의 유창성 점수 판단 및 보고 내용

과제 완성	유창성	보고 내용
일반적인 채점 방식 또는 전략에 대한 응답		일단은 과제 완성이 명확하게 언급되었는지 부분에서 개수가 counting되니까, 그 부분에서 유창성이랑 같이 가게 되죠. 그리고 그 특이한 경우는, 2, 3번 같은 경우를 한꺼번에 발음, 그러니까 문장으로 이어서 슬쩍 넘어가는 아이들이 있어요. 과제 점수는 낮지만, 유창성 부분에서는, 전체적인 흐름을 봤죠. 사실 여기서 2번과 3번에서, 3번을 언급하는 아이들이 많지 않았어요. 그니까, 2번 같은 경우, 엄마가 죘는데 거절했다. 한 번 더 말씀하시거든요. 그니까, 헬멧없인 할 수 없다라는 개념인데, 이 그림이 아이들한테 약간 혼란을 주는 것 같아요. 그래서, 그냥 엄마가 죘는데 얘가 거절을 했는데, 어쩔 수 없이 그냥 썼다, 이렇게 가요.
3	3	3점을 넘을 수 없을 것 같아요. 일단 과제 완성의 영향을 받게 되구요. 그래도 첫 번째 문장과 두 번째 문장에서는 자연스럽게 왔구요.
5	4	이거는 다소 pause가 막 들어가는 것 같아요.

채점자 SH-4는 표 6에서 보듯이 일반적인 채점 방식 또는 전략에 대한 응답에서 유창성에 대해 '억양(intonation)'과 '휴지(pause)', 그리고 "목소리의 톤(tone)"에 대한 판단으로 정의하고 있다. 이는 NEAT의 채점 기준에서 유창성을 정의한 것과는 차이가 있는데 억양은 발음의 구인(construct)으로 정의되어 있고 목소리 톤은 어느 영역 기술에도 없지만 채점자가 의미하는 바가 '전달하려는 내용이 얼마나 잘 들리도록 말하는가'에 관한 것이라면 역시 발음 영역의 기술인 "Rythm, intonation, and pronunciation are reasonably intelligible for delivering meaning."과 가장 관련이 있는 것으로 보인다. 이와 같이 이 채점자는 유창성을 발음과 혼동해서 정의하고 있으며 점수 판단 근거를 설명하는 데 계속 '(한국식) 발음'과 '억양', 이 두 가지로 유창성을 판단한다. 또한, 과제완성의 수준에 맞춰 유창성의 정도를 상대적으로 파악하고 있어 과제와 같

은 점수를 줄 경우에는 긍정적으로 평가하고 과제보다 1점을 낮게 줄 것이면 부정적으로 평가하는 양상을 보인다. 즉, SH-4도 대부분의 다른 상위 채점자들처럼 유창성 자체의 절대적인 판단 기준은 적용하지 못하고 있다. 또한, 유창성을 판단하는 데 영향을 주는 외적 요인으로 NEAT의 급수를 언급하였다. 같은 수준의 수행에 대해 2급보다 학력 수준이 낮은 학생들이 응시하는 3급 시험에서 더 후한 판단을 한다고 했다. SH-4는 0.7524로 우수 채점자 중에서는 그리 높지 않은 신뢰도를 보인 채점자인데, 채점의 일관성과는 별개로 유창성 판단의 타당성이 매우 의심스러운 경우이다. 발음과 유창성을 혼돈해서 기술하는 것을 보건데, 유창성 개념보다는 발음 개념으로 유창성을 이해하고 과제완성 수준에 맞춰 최종 점수를 결정하는 것으로 보이며 그 외에 유창성(또는 발음) 개념과 전혀 관계없는 NEAT의 급수라는 외적 요인을 점수 판단 기준으로 명시하고 있다.

표 6. SH-4의 유창성 점수 판단 및 보고 내용

과제 완성	유창성	보고 내용
일반적인 채점 방식 또는 전략에 대한 응답		fluency는 intonation도 보고, 그 학생이 pause가 얼마나 긴가 짧은가... 이런 것을 보는 거죠. 그러니까 fluency에서 문법은 제외가 되는 거죠. 일단은. 얼마나 의사소통을 pause라든지 intonation이라든지 그런 것들이 좋은가. 또 목소리의 tone이라든지... 그런 것을 보는 거죠.
5	4	발음에서는 intonation이 좀 약했고, 그리고 의사 전달하는 면에 있어서 거의 한국식 발음을 했단 말이에요. 그것이 완벽한 fluency를 형성하지 못했어요. pause같은 경우도 조금 끊어 읽기가 부족했고
4	3	아까 그 학생보다 훨씬 좀 떨어지는 거죠. 그러니까 발음 자체도 intonation이 아까 그 학생보다 약간 좋기는 하지만 pause라든지 끊어 읽기 등에서 약간의 문제점이 조금 보인다고 생각했어요.
3	3	(과제완성) 3점 선에서 볼 때, fluency같은 경우 이 학생은 다른 학생에 비해서 충분히 3점을 받을 수 있는 거죠. 아까 4점이나 3점을 받은 학생을 보면...
5	5	이 학생의 발음 조금 좀 억세긴 하지만 의사소통하는 데는 큰 문제가 없다는 거죠. 이런 경우도 마찬가지로 사실은 4.5점이예요. 딱 듣는 순간 4.5인데 없으니깐 4점을 줄지, 5점을 줄지 순간 되게 망설여져요. 그런데 이런 경우는 웬만하면 3급이니까... 2급의 경우는 조금 더 strict하게 하지만 3급의 경우는 4.5점이냐, 5점이냐가 될 경우에는 5점 정도로 주는 게...

채점자 SH-5는 채점이 가장 어려운 영역으로 유창성을 꼽은 이유가 과제완성 수준에 따라 유창성을 판단해야 하는지 아닌지가 채점자마다 관점이 다른데 자신은 급수에 따라 달리 적용한다고 하였다. 즉, SH-4와 마찬가지로 채점 영역 외에 NEAT 급수에 따라 유창성 점수 판단을 달리 한다고 하였다. 이 채점자는 “발화의 머뭇거림”, “지속적인 흐름”, “휴지” 와 같은 개념으로 유창성을 설명하는 데, 이는 NEAT의 유창성의 주요 개념 중 주로 pause에 초점을 둔 이해로 보인다. 골든룰 적용 후 점수를 상세하게 판단해야 할 때는 하고 싶은 말을 휴지 없이 죽 이어가는지 아닌지를 기준으로 점수를 결정한다. 표 7에 제시한 예들 중 두 가지는 머뭇거림 혹은 휴지가 별로 없어서 과제완성에 따른 상한 점수를 주는 것으로 쉽게 결정을 한 반면, 과제완성이 3인 예에서는 휴지가 있는 발화에 대해 2점을 줄 것인지 3점을 줄 것인지 제대로 결정을 못하고 있

다. 즉, 휴지가 없는 흐름에 대해서는 골든룰에 따른 점수 결정을 쉽게 하지만 판단이 어려운 경우에는 어느 정도의 휴지는 3점이고 어느 정도는 2점인지 적용할 분명한 기준이 내재화되어 있지 않은 것으로 보인다. 이 채점자의 예에서 보듯이, 과제완성이 정한 상한 점수 때문에 머뭇거림이나 휴지가 별로 없는 유창한 답안이라고 판단해도 최고 점수를 줄 수 없는 것은 골든룰 때문에 채점자가 유창성 결정을 타당하게 하지 못하는 사례라고 할 수 있다.

표 7. SH-5의 유창성 점수 판단 및 보고 내용

과제 완성	유창성	보고 내용
일반적인 채점 방식 및 전략에 대한 응답		태스크(과제완성) 같은 경우에는 어느 정도 명확한 기준을, 이거는 '뭐 몇 개를 이야기 했을 때 뭐 몇 개에서 할 수 있다' 이러는 데, fluency라는 개념이 태스크를 한 기준에서 어느 정도 fluency한 것인지, 아니면 전반적인 language나 이런 걸 발음, 모든 걸 포괄해서 해야 되는 건지에 있어서, 보시는 관점이 조금 다를 수 있는 것 같아요. 저는, 어... 급수에 따라 다른 편인데, 제가 이제 2급을 채점하느냐 3급을 채점하느냐에 따라 다른데, 2급에 있어서는 조금 fluency 부분이나 이런 부분에 조금 더 점수를 갖다가 조금 짜게 주는 편이고, 3급 같은 경우에는 되도록 후하게 주려고 마음을 먹고 있거든요. <중략>
4	4	그러니깐 이제 태스크를 4점을 줬기 때문에, 그런데 자기가 했던 답변에 있어서는 그렇게 머뭇거림은 없었거든요. 그래서 4점을 줬고.
4	4	유창성 부분에서 3점을 줄까 상당히 사실은 고민을 했었는데 이 아이 같은 경우에는 지가 말할, 목소리가 일단 조금 작은 면이 없지 않아 있어 가지고 인지하기가 조금 어렵긴 했지만 지속적으로 자기 나름은 휴지가 없이 한다고 주욱 이렇게 간 느낌이 많이 들었고요.
3	2/3	태스크가 3점임에도 불구하고 그 말을 하는데서 휴지가 조금씩 있었는데 아마 제가 채점을 한다면 몇 번 더 들어봐야 할 것 같다는 느낌이 들어요. 왜냐면 2점을 주기에는 약간 가혹한 면이 없지 않아 있는 것 같아서, 조금, fluency가 3점일까 2점일까 조금 몇 번 들으면서 고민을 할 것 같아요.

마지막으로, 표 8에서 보듯이 채점자 SH-6은 유창성을 "자연스럽게 발화가 쭈욱" 이어지는 것, "말이 자연스럽게 나와"서 pause가 없는 것으로 이해하는 것 같다. 따라서 "content word를 생각"하느

라 시간이 걸리는 것은 pause가 생기는 것이라 유창성이 부족한 것으로 판단하지만 이런 경우를 4점으로 할지 3점으로 할지에 대한 내재된 기준이 없기 때문에 판단을 망설인다. 다른 예에서도 “만점을 주기에 좀 부족”하다는 식으로 명확한 기준 없이 모호하게 점수를 판단하려고 한다. 즉, 전반적으로 pause라는 개념으로 유창성을 이해하지만 점수대별 판단 기준은 내재화하지 못하고 발화의 양과 같은 모호한 기준으로 점수를 정한다. 특히, 과제완성이 5점 만점이라 유창성에 상한선이 따로 없는 경우에도 “발화양이, 너무 작구요”라면서 순전히 발화의 양을 기준으로 유창성 점수를 결정했다.

표 8. SH-6의 유창성 점수 판단 및 보고 내용

과제 완성	유창성	보고 내용
4	4	음, 아주 능숙하지는 않지만, 발음을 자연스럽게 어느 정도는 했다고 생각했기 때문에
5	4	(그림)1번하고, 1번은 뭐 완벽하게 됐지만, (그림)2번 같은 경우도 전체적으로 발화가 자연스럽게 쭈욱 나오기보다는 어떤 content word를 생각하는 게 좀 많이 시간이 걸린 것 같구요. 그래서 인제 5점 주기보다는 한 4점 정도? 그렇지만 자기 하고 싶은 얘기를 쭉 했다는 게 있어서 3점 주기는 좀 야박한 것 같아서 4점을 줬고
5	4	만점을 주기에는 좀 부족한 면이, (그림)1번에서 발화(양)가 조금 부족한 면이 사실 있었고, 그 다음에 뭐, (그림)3번 같은 경우도 말을 하면서 자연스럽게 나왔다기 보다는 좀 pause가 군데군데 많이 보여서
5	3	이 학생같은 경우는, 3점을 주고 싶어요. 발화양이, 너무 작구요, 필요한 정보는 말했지만 너무 작은 게 있고,
3	2	자연스럽게 발화되는 게 거의 없는 것 같아요. 그렇다고 1점 줄려니까 발화 내용이 어느 정도 확보됐기 때문에 2점 정도를 제가 준 것 같구요.

종합하면, 6명의 상위 수준 채점자들 중 SH-4를 제외한 다섯 명의 채점자는 유창성에 대해 ‘휴지(pause)’, ‘머뭇거림’, ‘흐름이 있는 자연스러운 발화’ 등으로 NEAT의 정의에 부합하는 이해를 보였다. 다만, SH-1을 제외하고는 ‘속도(rate of speech)’라는 개념을 거의 언급하지 않았는데, 이는 아마도 채점자들 스스로가 비원어민인 영어교사로서 우리나라 고등학생들이 속도감 있게 영어를 말하는 것보다는 끊어짐 없이 말을 이어갈 수 있는 능력을 더 중요하다고 판

단하기 때문이 아닐까 추측해본다. SH-4는 적어도 자신이 한 설명에 근거해 볼 때 유창성의 개념을 제대로 이해하지 못하고 발음의 개념과 혼동하고 있는 것 같다. 점수대별 판단 기준에 대해서는 SH-2만 유일하게 절대적 기준을 내재화하여 적용하고 있는 것으로 보이고 나머지 채점자들은 과제완성 수준에 따라 상대적인 판단을 하거나 전체 발화의 양을 고려하거나 NEAT의 급수를 고려하는 등 유창성 자체의 기준이 아니라 외부 조건을 적용하는 것으로 보인다.

결국, SH-2와 같은 일부 채점자를 제외하고 대부분의 채점자들이 유창성 개념 자체에 대한 이해는 어느 정도 타당성이 있으나 점수대별 판단 기준을 내재화하여 적용하지는 못하였다. 이는 NEAT에서 제공하는 유창성의 점수대별 판단기준을 적용하기 쉽지 않기 때문일 수도 있지만, 한편으로는 NEAT와 같은 대규모 말하기 시험에서 채점의 일관성을 높이기 위해 채택한 골든룰에 의존을 하게 되면서 채점자 스스로 유창성에 대한 이해를 근거로 점수대별 판단 기준을 내재화할 필요성을 크게 느끼지 못하게 된 것으로도 보인다. 실제 일부 참여자들은 면담 도중 자신이 과제완성 점수를 기준으로 타 영역 점수를 얼마나 기계적으로 빠르게 결정할 수 있는 지 강조하기도 하였다.

4.2 채점자들의 발음 이해 및 점수 판단 기준

NEAT에서 발음은 '자음과 모음의 발음', '단어와 문장 강세(stress)', '억양(intonation)', '연음(sound-linking)' 등의 개념으로 정의하고 점수대별 판단 기준은 '의사소통을 방해하는 정도'로 구분하여 다소 복잡해 보인다. 하지만 실상은 구성 영역과 더불어 3점 척도를 사용하고 과제완성에 따른 골든룰의 제약 없이 '알아들을 수 있는 정도(intelligibility)'를 가장 중요한 기준으로 하고 있기 때문에 5점 척도를 사용하는 유창성에 비해 비교적 간단하다고 할 수 있다. 다만, 발음 점수 판단에는 골든룰이 적용되지 않음에도 불구하고 아래 논의에서 보듯이 많은 채점자들이 과제완성 영역 점수를 고려하는 것으로 보인다. 따라서 이후 예시 자료에도 발음 점수와 함께 과제완성 점수도 함께 제시한다.

표 9에서 보듯이, SH-1은 다른 영역에 비해 발음은 별 다른 설명이 필요 없을 만큼 점수가 분명하다고 여기는 듯하고 결과적으로 다른 채점자들보다 발음에 관한 설명이 간단하다. 하지만 이 채점자가 발음 점수를 쉽게 판단하는 것은 과제완성 영역과는 상관없이 독립적으로 판단을 해야 하는데도 골든룰을 점수 판단 근거로 포함시키기 때문인 것으로 보인다. 즉, 과제완성 점수가 5점 중 3점인 경우 "발음은 상관없이" 최고점 3점 대신 2점을 부여한다. 물론 "뭐 이 정도면 1점보다 높구요"의 경우와 같이 발음 자체의 '이해가능성' 기준을 적용하기도 한다. 결국, 이 채점자는 NEAT의 발음 영역 주요 개념들을 거의 고려하지 않고 단순하게 '이해가능성'과 과제완성 점수를 판단 근거로 사용하는 것 같다.

표 9. SH-1의 발음 점수 판단 및 보고 내용

과제완성	발음	보고 내용
3	2	과제완성이 (5점 중) 3점이면 발음은 상관없이 2점
2	2	발음은 뭐 이정도면 1점보다 높구요.
5	2	약간 어색한 게 몇 개 있었죠.

표 10에서 보듯이 채점자 SH-2는 '자음과 모음의 구분', '음소의 구분', '의사소통에 대한 방해', '이해가능성(intelligibility)', 등의 개념으로 발음을 설명하고 있는데, 이는 NEAT에서 정의한 발음의 주요 개념들과 거의 일치한다. 특히, 이 채점자는 자음과 모음의 발음이 정확한가, 의사소통을 방해하는 틀린 발음이 있는가의 여부를 두고 3점과 2점을 구분하고 있다. 또한 실제 예는 아니지만 2점과 1점을 구분하는 기준으로 대체적으로 음소를 구분해서 발음할 수 있는지의 여부로 보고 있다. 앞서 유창성 영역에서도 SH-2 채점자가 유일하게 점수대별 판단 기준을 내재화하고 있었는데 역시 발음에서도 비교적 점수대별 분명한 판단 기준을 적용하는 것 같다.

표 10. SH-2의 발음 점수 판단 및 보고 내용

과제완성	발음	보고 내용
2	2	발음 같은 경우 1점을 받는 아이들이 그렇게 많지를 않더라고요. 제가 채점을 해 보니깐 웬만하면 그럭저럭, 아주 정확하지는 않다 하더라도 자음 모음 구분할 정도로 아이들 수준이 되기 때문에 대체적으로 이제 2점에 들어가고
4	3	뭐 이렇게 크게 문제될 만한, 커뮤니케이션을 해칠 만큼 크게 문제가 되는 음소 변경이라든지 이런 것들이 눈에 띄지 않아 3점을 줬고
4	2	이 아이는 문제가 friend라든지 food라든지 뭐 이런 데서 끝에 인제 자음처리가 그렇다고, 또, 또 한 가지는 약간 intelligibility가 조금 떨어진다고 봐서 만점을 안 줬고요.

표 11에서 보듯이, 채점자 SH-3은 발음 점수 판단에 발화의 이해가능성 기준을 유일하게 적용하고 그것도 가급적 관대하게 적용하는 것으로 보인다. 일반적인 채점 기준 및 전략에 관한 답변을 보면 사투리 예를 들면서 발음이 좀 이상하더라도 이해가 되면 3점이라서 그 빈도가 가장 높다고 한다. 실제 채점에서도 자신이 “들을 수 있다”는 기준으로 3점을 부여한다. 하지만 어떤 경우에 2점을 주는 지는 예를 찾지 못했다. 결국 아예 이해하지 못할 정도의 경우를 제외하고는 대부분 3점을 많이 주다보니 2점의 예를 찾기 어려웠을 것으로 보인다. 즉, 이해가능성 기준을 너무 단순하게 적용하여 채점에서 타당성이 다소 부족한 것으로 보인다.

표 11. SH-3의 발음 점수 판단 및 보고 내용

과제완성	발음	보고 내용
일반적인 채점 방식 및 전략에 대한 응답		굉장히 발화양이 적을 때 1점이고, 웬만한 경우는 다 2점이에요. 그리고, 음, 우리 식으로 발음하더라도 유창하게 가는 경우라면 이 발음에서도 저희 그 왜, 음, 외국인들도요, 사투리라던가 아님 영국식 영어, 음, 호주 영어가 있듯이, 그래도 이해가 된다면, 그 부분에서는 발음 점수가 3점이 나오게 되더라구요. 그래서 음, 3점이 많이 나오게 되요. 많이 나오는데, 쫌 이상한 게 몇 번 나오는 게 있죠. 3점은, 빈도수로 보면 오히려 3점이 더 많은 것 같아요. 웬만큼 응답을 하면.
2	3	사실은 좀 더 깎고 싶긴 하지만, 제가 들을 수 있다는 얘기는 전달이 된 거고.
3	3	그렇게 틀리지는 않았으나, 3점을 주기에는 참 어려운 부분이긴 한데, 원래 연수상으로는 독립 채점을 하라 했어요. 어제 연수에서는 다시 그랬는데, 독립 채점을 한다면 3점을 줄 수 있을 것 같아요.

SH-4는 '억양(intonation)'을 언급하기는 했지만 실제 점수 판단 기준으로 활용한 것으로 보이지는 않았다. 표 12에서 보듯이 발음에 한국식 악센트가 있어도 알아들을 수 있으면 3점을 주지만 2점을 주는 경우는 '발음이 좋지 않다' 또는 '알아는 듣는다' 정도로 기술하고 있다. 발음이 좋지 않다는 것을 구체적으로 설명하지 않고 "어쨌든 알아들을 수 있는 데 지장이 없으니까"라고 하면서 3점을 준 답안도 있고 "알아는 들으니까 이 정도"라고 하면서 2점을 준 경우도 있어 그 기준이 무엇인지 잘 드러나지 않는다. 하지만 "들어보시다시피"라고 말하는 것으로 보아 그 구분이 본인에게는 꽤나 명확한 것으로 보인다. 어쨌든 이 채점자가 적용하는 유일한 발음 판단 기준은 이해가능성으로 보이고 그 정도를 나름 구분하고 있는 것으로 보인다.

표 12. SH-4의 발음 점수 판단 및 보고 내용

과제완성	발음	보고 내용
5	3	intonation이 좀 약했고, 그리고 의사 전달하는 면에 있어서 거의 한국식 발음을 했단 말이에요. 이 정도 발음은 우리가 알아들을 정도여서 의사소통에 문제가 없다고 해서
4	3	어쨌든 들을 수 있는 데 지장이 없으니까
3	2	그 대신에 발음이 그렇게 좋지는 않잖아요. 들어보시다시피... 그래서 1점을 깎은 거고...
5	2	알아는 들으니까 이 정도...그러니까 발음이 보통 2점

SH-5는 표 13에서 보듯이 발음 점수를 결정하는 데 과제완성 점수를 절대적으로 고려하였다. 과제완성이 5점이면 발음도 거의 만점, 과제완성이 4점이면 발음은 3점이나 2점, 그렇지만 과제완성이 그 이하라도 2점보다 더 낮게 주는 경우는 드물다고 하였다. 과제완성이 5인 마지막 경우를 보면 과제완성에 따라서 같은 발음을 3점 또는 2점으로 달리 채점할 수 있다고 하여 이 채점자가 발음 영역 판단에 과제완성을 언제나 고려하는 것을 알 수 있다. 이 채점자는 과제완성이 4점인 경우 발음을 3점 또는 2점으로 결정해야 하기 때문에 비로소 intonation과 같은 발음 고유의 개념을 고려하는 것으로 보인다. 과제완성을 기준으로 발음 점수를 정하다 보니 다른 채점자들과는 달리 이해가능성이 주요 판단 기준으로 언급된 예가 없다. 또 하나 특이한 점은 발음 영역과 구성 영역을 그냥 함께 묶어서 “발음과 구성력은 조금 약했고”처럼 함께 판단하는 경향이 있었다. 표 13에 제시된 예 이외에도 이런 경우가 몇 차례 더 있었다. 아마도 이 채점자는 아주 명백하게 차이가 나는 경우를 제외하고는 과제완성을 기준으로 3점 척도를 사용하는 이 두 영역의 점수를 동일하게 판단했을 것으로 보인다. 결국, 이 채점자는 발음에 대한 자신의 이해를 적용해서 점수를 판단하는 것이 아니라 과제완성에 종속시켜 점수의 범위를 축소시키는 방식의 채점을 하여 그 타당성이 크게 의심스러운 사례이다.

표 13. SH-5의 발음 점수 판단 및 보고 내용

과제완성	발음	보고 내용
일반적인 채점 방식 및 전략에 대한 응답		사실 발음이 1,2,3점으로 되어 있어...발음도 마찬가지고 Discourse(구성)부분도 마찬가진데, 그게 3점으로 되면서 애매한 부분의 의미가 그렇게 크게 있나? 굳이 점수를 그렇게 매겨야 하나 싶은 생각이 가끔씩 있어요. 그니깐 그냥 전반적으로 태스크하고 영향을 많이 받는 편인데, 태스크가 5점 3점으로 하면, 그냥 3점 쪽에 왔다갔다 하고, 그런데 이제 태스크가 한 4점, 이렇게 가더라도. 발음 부분이 조금 약하면 2점을 주는데, 나머지 432 이렇게 가면, 그렇다고 1점을 주기에는 조금 그런 면이 있으니까, 대부분은 2점 선을 넘어가지 못하는 경향이 많이 있어서, 그 부분의 점수를 조금 더 세분화 하던지, 아니면 아예 그게 무슨 큰 의미가 있나? 이런 생각이 전 좀 많이 들더라구요.
4	2	발음을 3점을 줄까 2점을 줄까 상당히 사실은 이걸 고민을 하던 파일이었는데 나름 발음이 뭐 인지는 알 수 있는데 아이가 intonation이 조금 없이 이렇게 쭉 가는 느낌이 조금 들어가지고 그랬고,
4	2	발음과 구성력은 조금 약했고
5	3	뭐 그 좀 깨끗하긴 했거든요. 전반적인 게 이게 이 아이가 만약 태스크가 조금 더 약하다면, 2점대로 내려갈 수도 있지 않을까 하는 생각이 들긴 하는데, 태스크를 다 했기 때문에 3점 주고

마지막으로, 표 14에서 채점자 SH-6이 발음 최고점수 3점을 준 이유를 기술하는 것을 통해 보건데, 이 채점자는 발음을 '음소 발음', '억양', '사고 단위(thought group)', '이해가능도'로 이해하고 있다고 할 수 있다. 이 중 사고 단위는 발음 보다는 유창성의 개념이지만 대체적으로 이 채점자의 발음에 대한 이해는 타당한 것으로 보인다. 하지만 이 채점자의 실제 채점은 이 발음에 대한 이해를 적용한 것으로 보기 어렵다. 첫 번째와 세 번째 예에서는 발화의 양을 기준으로 3점 대신 2점을 주었다. 두 번째 예에서는 "자연스럽게 흘러가서"라고 하여 오히려 유창성이 판단기준으로 사용되었다. 마지막 예는 설명은 길지만 결국 과제완성이 5점이므로 부족한 부분이 있어도 발음을 3점으로 정했다. 첫 번째와 세 번째 발화의 양을 기준으로 했던 예에서도 사실 발화의 양이 과제완성과 관련이 있다는 것을 고려하면 이 두 경우도 결국 과제완성이 주요 판단 기준이었을 것으로 보인다. 그리고 마지막 예에서 언급했듯이 NEAT

급수도 발음 점수 판단에 영향을 주었다. 결국, 이 채점자는 자신이 발음의 개념을 이해하고 있는 것과는 별개로 실제 채점에서는 과제 완성, 발화의 양, NEAT 급수 등 발음과는 별개의 기준들을 적용해서 점수를 결정하는 것으로 보인다.

표 14. SH-6의 발음 점수 판단 및 보고 내용

과제완성	발음	보고 내용
3	2	발음이 나쁘다기보다는 발음을 평가할 수 있는 근거가 좀 많이 부족한 점이 있기 때문에
4	3	들어봤을 때 크게 거슬리는 게 없고 자연스럽게 흘러가서
4	2	두 번째 발화, 문항을 전혀 대답을 못했기 때문에 3점 받기는 좀 부족하다고 생각을 한 것 같구요.
5	3	제가 한 번 들었는데 큰 어려움 없이 들을 수 있었고, 음소 발음이라든지, 그 thought group으로 나눠지는 것도 자연스러운 것 같애요
5	3	엄밀히 말하면, 3점 척도니까 그렇지만, 사실은 2.5점이나 2점 초반에 가까운 게 더 가까운 생각이 들어서, 발화양이 많진 않아서, 좀 그렇지만, 과제 완성이 이런 경우 5점으로 들어갔기 때문에, 3점으로 제가 주고 싶습니다. 웬만하면 3, 3급에 1번 같은 경우는 가장 쉬운 문항에 속하기 때문에 되도록 학생들한테 쫌 이래 자극을 주기보다는 장려한다는 측면에서 줬구요.

요약하면, 6명의 채점자들 중 2명을 제외하고는 NEAT에서 정의한 발음의 개념들을 이해하고 있는지 파악하기 어려웠는데, 이는 원어민 채점자들과는 달리 내국인 채점자들이 발음의 주요 개념들을 거의 언급하지 않았던 Lee 외 2인(2014)의 연구 결과와 크게 다르지 않다. 아마도 이해가능성을 주요 채점기준으로 적용하기 때문에 다른 세부 개념들을 고려할 필요가 없었을 수도 있다. 하지만 SH-2는 발음의 주요 개념들을 점수대별 판단 기준으로 적용하는 이상적인 행동을 보였다. SH-3은 이해가능성 기준을 너무 관대하게 적용하여 최고점을 남발하는 양상을 보였다. 또한 다수의 채점자들이 이해가능성 기준에 앞서 과제완성 영역 점수를 고려하는 양상을 보였다. NEAT에서 발음이 골든룰의 적용을 받지 않는 독립적 채점이 요구되는 영역임에도 불구하고 채점자들이 과제완성 점수를 고려해서 발음 점수를 조정하는 등 채점의 타당성이 결여된

행동을 보였다. 그 밖에도 발음 외적 요소인 발화의 양, NEAT의 급수 등을 고려하는 채점자도 있었다. 결국, 다수의 채점자들이 발음 영역 점수를 결정하는 데 '이해가능성'이라는 단순한 기준을 적용했고 그 수준도 절대적인 기준을 적용해 판단하기 보다는 과제완성 영역 점수를 고려하면서 조정하는 것으로 보였다.

5. 결론 및 시사점

본 연구는 우리나라 영어 학습자를 대상으로 하는 말하기 시험에서 분석적 채점 방식을 채택할 때 비원어민 채점자들이 유창성과 발음을 어떤 개념으로 이해하고 어떤 기준을 적용해서 점수를 판단하는지 분석함으로써 채점자들이 얼마나 타당성 있게 채점을 하는지 파악하고자 하였다. 그 결과를 요약하면, 본 연구 참여자들 대부분이 NEAT에서 정의한 유창성의 주요 개념들을 이해하고 있었고 구술 보고에서 활용하였다. 하지만 실제 채점에서는 그러한 개념들의 절대적인 수준을 판단해서 점수를 정하기보다는 과제완성, 발화의 양, NEAT 급수 등 유창성과 관계없는 외적 요인을 고려하였고 특히 과제완성 점수를 기준으로 유창성의 수준을 상대적으로 판단하였다. 발음 영역에서는 참여자들 대부분이 NEAT에서 정의한 발음의 주요 개념들을 고려하지 않았다. 대신 '이해가능성'을 발음 영역의 유일한 주요 판단 기준으로 적용했는데 이마저도 구체적인 점수 판단을 위한 절대적인 기준이 없었고 너무 관대하게 적용하는 모습을 보였다. 또한, 다수의 채점자들이 유창성과 마찬가지로 여전히 과제완성 점수를 고려해 발음 점수를 조정하기도 하였다. 결국, 본 연구 참여자들이 신뢰도가 높고 NEAT 채점 경험이 비교적 풍부한 채점자들임에도 불구하고 구체적인 채점 기준 적용에서는 타당성을 제대로 확보하고 있다고 보기 어려웠다. 즉, NEAT 특유의 골든룰을 능숙하게 적용해서 일관성을 유지하며 채점을 하는 능력과는 별개로 각 영역의 구인을 제대로 이해하고 점수대별 절대적인 판단 기준을 내재화하여 적용하는 능력은 부족하다고 할 수 있다. 여섯 명의 연구 참자가 중 SH-2만이 채점의 일관성과 타당성을 모두 확보한 것으로 보이는 유일한 채점자였다.

본 연구의 결과를 근거로 다음의 시사점을 논할 수 있다. 첫째, 말하기 채점의 타당성은 일관성만큼이나 중요한 요소임에도 불구하고 NEAT와 같은 공교육 내에서 행해지는 대규모 평가에서는 논란의 여지가 큰 채점의 신뢰도에 비해 덜 강조되는 것이 사실이다. 하지만 학생들에게 채점 영역별 장·단점에 대해 피드백을 제공하는 교육적 목적을 제대로 수행하려면 채점의 타당성이 확보되어야 점수가 의미하는 절대적인 수준을 알려줄 수가 있다. 향후 NEAT와 같은 대규모 말하기 평가를 실시한다면 이전의 경험을 거울삼아 골든룰과 같은 일관성 향상을 위한 전략만이 아니라 영역별 채점 루브릭을 보다 적용하기 쉽게 만드는 것을 포함해 채점의 타당성 확보를 위한 전략을 고민해야 할 것이다. 본 연구에서 보듯이 어떤 채점자는 골든룰을 적용해서 채점의 일관성 및 속도는 향상되지만 이 전략으로 인해 채점의 타당성이 위협을 받기도 하였다.

둘째, 본 연구 참여자들 대부분이 NEAT의 유창성 정의에 포함되는 'rate of speech', 즉 발화의 속도에 대해 거의 언급하지 않았다. 이는 말하기 지도와 평가가 활발히 이뤄지지 않는 우리나라 상황을 고려해 채점자들 스스로 발화의 속도보다는 주로 휴지와 머뭇거림의 정도로 유창성 수준을 판단하는 것으로 볼 수 있다. 하지만 선행연구에서 보았듯이 'speech rate'은 가장 중요한 유창성 변인이기 때문에 우리나라 상황을 고려해 어느 정도로 중요하게 평가되어야 할지 절충점을 찾는 방안에 대한 논의가 필요할 것이고 이것과는 별개로 채점자 연수에서 강조되어야 할 개념이다.

셋째, 채점자들이 발음 채점에서 적용되는 주요 기준인 이해가능성의 개념을 좀 더 구체적으로 이해하도록 안내해야 할 것이다. 국제어로서 영어의 다양한 발음을 존중하여 EFL 상황의 우리나라 학생들이 원어민과 같은 발음이 아니더라도 이해할 수 있는 정도의 발음이면 된다는 입장을 반영한 기준이 이해가능성이고 이 취지는 충분히 공감할만하다(Jenkins 2000). 하지만 본 연구에서 보듯이 이는 발음에 대해 지나치게 관대한 평가를 하는 근거가 되어 변별력 없이 적용될 가능성이 크다. 이 개념을 '무슨 말을 하는 지 알아들을 수 있다'라고 단순하게 적용할 것이 아니라 한국인 악센트가 있더라도 자음과 모음을 제대로 발음하고 적절한 억양을 사용하는 능력으로 구체적인 기준을 적용하도록 하는 채점자 안내가 필요하

다.

마지막으로, 본 연구 참가자 중 SH-2는 소위 이상적인 채점자로 보인다. 이 채점자는 NEAT의 채점 루브릭을 거의 외우고 있었고 그냥 말로만 그 문구를 가져다 쓰는 것이 아니라 실제 그 내용을 적용해서 채점을 하였다. 다른 참가자들도 똑같은 연수를 받고 채점 참여 횟수도 비슷하지만 이렇게 루브릭에 충실한 채점자는 찾지 못했다. Davis(2016)의 연구에서도 가장 정확한 채점자는 제공된 점수대별 예시답안을 더 자주 참조한다고 하는데 이는 루브릭에 충실하다는 것과 통한다. 결국, 채점자 개인의 능력 또는 노력도 중요하지만 루브릭을 잘 적용한 예시답안과 함께 반복적인 연습 기회를 제공하는 연수를 통해 좋은 채점자를 확보할 수 있을 것이다. 비록 현재 실상은 대규모 말하기 평가에 대한 관심과 논의가 활발하지 않고 채점자 연수도 거의 없는 상황이지만, 향후 교사의 말하기 지도와 평가 전문성 향상을 위한 연수 프로그램을 구성하고 시행하는 데 본 연구의 결과가 활용될 수 있기를 기대해본다.

참고문헌

Bosker, H., A. Pinget, H. Quene, T. Sanders and N. deJong. 2013. What makes speech sound fluent: The contribution of pauses, speed and repairs. *Language Testing* 30(2), 159-175.

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

Brown, A. 2005. *Interviewer Variability in Oral Proficiency Interviews*. Frankfurt, Germany: Peter Lang.

Campione, E. and J. Véronis. 2002. A large-scale multilingual study of silent pause duration. Paper presented at the Speech Prosody Conference, Aix-en-Provence, France.

Chalhoub-Deville, M. 1995. Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12(1), 16-33.

Davis, L. 2016. The influence of training and experience on rater performance in scoring spoken language. *Language Testing* 33(1), 117-135.

Derwing, T., M. Rossiter, M. Munro and R. Thomson. 2004. Second language fluency: Judgments on different tasks. *Language Learning* 54(4), 655-679.

Ginther, A., S. Dimova and R. Yang. 2010. Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing* 27(3), 379-399.

Jenkins, J. 2000. *The Phonology of English as an International Language*. Oxford: Oxford University Press.

Kim, H. J. 2010. An investigation of novice, developing, and experienced raters' rating patterns on a second language speaking assessment. K*orean Journal of Applied Linguistics* 26(4), 151-182.

Kim, Y. 2009. An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing* 26(2), 187-217.

Kormos, J. and M. Denes. 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System* 32, 145-164.

Lee, J. O., H. W. Lim and H. J. Kim. 2014. An investigation into native English speaking and Korean raters' judgments of Korean English learners' pronunciations. *Modern English Education* 15(1), 195-216.

Meiron, B. E. and L. Schick. 2000. Ratings, raters, and test performance: An exploratory study. In A. J. Kunnan ed., *Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida*, 153-176. Cambridge: Cambridge University Press.

Merion, B. E. 1998. Rating oral proficiency tests: A triangulated study of rater thought processes. Paper presented at the Language Testing Research Colloquium, Monterey, CA.

Orr, M. 2002. The FCE speaking test: Using rater reports to help interpret test scores. *System* 30(2), 143-154.

Papajohn, D. 2002. Concept mapping for rater training. *TESOL Quarterly* 36(2), 219-233.

Raupach, M. 1987. Procedural learning in advanced learners of a foreign language. In J. A. Coleman and R. Towell, eds., *The Advanced Language Learner*, 123-155. London: CILT.

Rossiter, M. J. 2009. Perceptions of L2 fluency by native and nonnative speakers of English. *Canadian Modern Language Review* 65(3), 395-412.

Song, M. Y. and Y. S. Lee. 2015. Scoring behaviors of English speaking raters: Suggestions for rater training. *Journal of Research in Curriculum & Instruction* 19(4), 1081-1101.

Yun, W. 2009a. Discrepancy between Korean and native English raters evaluating the English pronunciation spoken by Korean learners of

English. *The Journal of Linguistic Science* 48, 201-217
Yun, W. 2009b. An analysis of the Korean inter-rater difference in evaluating English pronunciations of Korean speakers. *Studies in Foreign Language Education* 23(2), 85-103.
Zhang, Y. and Elder, C. 2011. Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing* 28(1), 31-50.

예시 언어(Examples in): 영어(English)
적용가능 언어(Applicable Languages): 영어(English)
적용가능 수준(Applicable Level): 대학 이상(Tertiary)

송민영
서울 종로구 북촌로 106
고려사이버대학교 실용외국어학과
전화: 02-6361-1863
E-mail: mysong88@cuk.edu

논문접수: 2017년 10월
논문수정: 2017년 11월
게재결정: 2017년 12월