

말하기 평가 방법간 신뢰도 분석 연구

— 초등학생의 설득하는 말하기를 중심으로 —

조재윤*

〈차 례〉

- I. 서론
- II. 말하기 평가의 기준 및 방법
- III. 연구 문제 및 연구 설계
- IV. 연구 결과
- V. 결론 및 제언

I. 서론

1. 연구의 필요성 및 목적

교육평가는 교육적인 성과를 얻기 위해 교육목표를 세우고 그 목표를 얼마나 달성했는지를 측정, 의사결정 또는 가치 판단을 하는 활동이다.

1900년대 초반 교육평가는 과학의 발달로 인하여 수치와 통계에 대한 믿음이 생겼으며 이로 인해 인간의 지적 능력이나 언어 능력, 교육적인 태도, 심리적 특성을 검사나 측정을 통하여 수치 또는 양적으로 나타내는 ‘양적 평가’에 대하여 신뢰를 갖게 되었다. 이러한 양적 평가의 대표적인 것이 표준화 평가이다.

* 서울교육대학교 강사.

표준화 평가는 오랜 시간, 많은 노력과 모의 검사 등을 통하기 때문에 평가의 신뢰성이 매우 높고, 객관적이며 학생의 선발과 분류를 하는 대규모 평가에 유용하였다. 그러나 표준화 평가와 같은 양적 평가는 행동주의 심리학에 기반을 두었기 때문에 학생들이 습득해야 할 지식이나 기능 등을 하위 요소로 나누어 측정하였으므로 학생의 학습 과정에 대한 이해, 학습 과정에 포함되는 문제해결력, 고등정신능력 등을 측정하는데 부적합하였다. 그래서 새로운 평가의 대안으로 구성주의에 기반을 둔 '질적 평가'가 논의되었다.

질적 평가는 신뢰성보다는 타당성을 중시하고 학습 과정을 파악하게 해 주며, 교육과정과 교수·학습 방법을 개선하는데 장점이 있다. 그리고 학생의 개인차를 존중하여 학생의 학습 과정에 대한 상세한 정보를 제공해 주는 장점이 있다. 그래서 현재의 평가 패러다임은 질적이면서 실제적 평가인 수행평가가 주류를 이루고 있다.

교육평가의 패러다임 변화 속에서 각 교과 교육에서도 교과 효과적 인 교육을 위해서 다양한 평가 시도가 이루어지면서 관련 연구가 이루어 지고 있다. 국어 교과에서도 1990년대 초부터 국어교육학의 평가 일반론과 각 하위 영역의 평가 방법론에 대한 논의가 많아졌다. 그런데 국어과의 하위 영역 중에서 읽기 영역과 쓰기 영역에 관련된 평가 연구는 많이 수행되었으나 말하기와 관련된 평가 연구는 매우 부족한 현실이다.

국어과 평가는 학생의 국어 능력을 제대로 평가할 수 있어야 한다. 그러나 표준화 평가와 같은 양적 평가로는 학생의 국어 능력을 제대로 평가할 수 없다. 읽기 평가의 경우, 객관식 위주의 표준화 평가의 상황에서의 읽기와 일상생활 장면에서의 읽기는 여러 점에서 차이 - 지문의 유형, 읽기 목적, 읽기 과정의 지식의 출처, 시험 상황에서 사용하는 측정 기능과 일반 읽기 상황에서 사용하는 읽기 기능 - 가 많기 때문에 객관식 위주의 표준화 평가로는 학생들의 실제 읽기 능력을 파악할 수 없으며, 평가의 타당도가 높지 않다.

읽기가 아닌 말하기의 경우는 어떠한가? 국어과의 말하기 평가를 객관식 위주의 표준화 평가를 실시할 때, 타당도는 어떠한가? 객관식 평가

나 지필 형식의 평가가 학생의 말하기 능력을 제대로 평가할 수 있는가? 말하기를 객관식이나 지필 형식의 방법으로 평가하는 것은 타당도가 매우 낮다고 할 수 있다. 지필 형식의 평가를 잘 치른 학생이 실제로 말하기를 잘한다고 보장할 수 없기 때문이다. 실제 말하기 능력이 어떠한지는 학습 상황이나 생활 장면에서 실제로 말하는 것을 평가해야 한다. 그래서 실제적이면서 질적인 수행평가를 통하여 말하기 능력을 알아보려고 한다.

그러나 말하기 평가 장면에서 발생하는 중요한 두 가지 문제가 있다. 먼저, 생태학적으로 실제 생활에서의 말하기와 같은 말하기 평가를 하는 것은 불가능하다. 다음으로, 말하기 평가의 신뢰도를 알기 어렵다. 첫 번째 문제점은 학습 장면을 될 수 있는 대로 생활 장면에 가깝게 하여 평가함으로써 어느 정도 해결할 수 있다. 하지만 근본적으로 교실 학습 장면에서 이 문제를 완전하게 해결하는 것은 불가능하다. 두 번째 문제점인 말하기 평가의 신뢰도의 문제는 충분히 해결할 수 있다. 학생들의 말하기 평가자는 학생들 간에 이루어지는 경우도 있지만 거의 대부분 교사들이다. 평가자들의 말하기 평가 방법의 신뢰도가 어떠한지는 통계적인 방법으로 파악할 수 있다. 평가자들의 평가 방법이 다른 총체적 평가와 분석적 평가의 비교를 통하여 신뢰도를 알아볼 수 있다.

이 연구는 초등학교 6학년 학생들이 수행한 설득하는 말하기에 대하여 교사 평가자들이 실시한 평가 방법에 따른 신뢰도의 상관관계를 통하여 말하기 평가 방법간 신뢰도를 알아보고자 한다. 이는 교사들이 총체적 평가와 분석적 평가를 실시할 때 평가자들이 어떤 점에 유의해야 하는지에 대한 시사점을 줄 뿐만 아니라 평가의 전문성을 높여야 한다는 인식을 갖게 할 것이다. 또한 평가자인 교사의 평가 전문성을 높이기 위한 정책적인 방안을 마련하는 기초가 될 것이다.

2. 연구의 제한점

이 연구는 다음과 같은 제한점이 있다.

첫째, 말하기 수행 평가자들의 사전 훈련이 이루어지지 않았다. 그러나 교육 현장에서 말하기나 쓰기 평가를 실시할 때 교사들의 평가자 훈련이 거의 이루어지지 않는다는 점을 생각하면 현실적인 교육 현장의 평가자 모습을 그대로 보여준다고 할 수 있다.

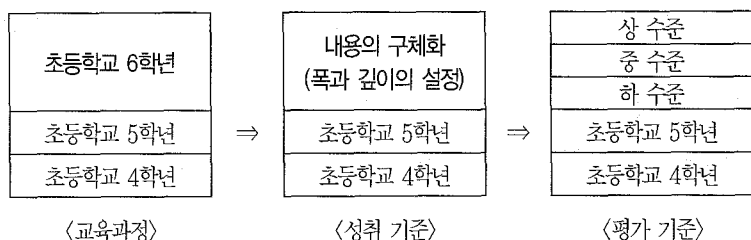
둘째, 평가자들은 캠코더로 촬영한 화면을 보고 평가하였다. 실제 말하기 평가는 말하기 과정에서 청자의 반응이나 태도 같은 말하기 상황이 고려되어야 하지만 이번 평가에서는 이러한 변수는 반영되지 않았다. 그러나 이는 교실 수업에서 이루어지는 학생의 수행 장면을 많은 평가자들이 동시에 평가할 수 없는 말하기의 특수 상황이라고 할 수 있다.

II. 말하기 평가의 기준 및 방법

1. 말하기 평가의 기준

말하기 평가에서 평가 목표를 수립하고 성취 기준을 선정하여, 평가 과제가 제시되고 나면, 그 다음으로 교사들에게 필요한 것은 평가 기준이다. 평가 기준(assessment standard)이란 평가 활동에서 실질적인 기준 역할을 할 수 있도록 각 평가 영역에 대하여 학생들이 성취한 정도를 몇 개의 수준(예: 상, 중, 하)으로 나누어, 각 수준에 기대되는 성취 정도를 구체적으로 진술한 것이다. 즉, 평가 기준이란 평가 목표에 도달하기 위하여 학습자가 취한 수행 활동이나 활동의 결과물에 대하여 판단하려는 기준이다. 평가 기준은 학생의 수행 능력을 질적으로 판단할 수 있는 정보를 지녀야 하고, 수준을 구별할 수 있는 근거가 명확해야 한다. 그러

므로 평가 기준은 평가 영역별로 학생들의 성취 정도의 단계를 보통 3단계 또는 4단계 정도로 판정하는데 실질적인 지침이 되도록 평가 영역별로 성취 기준의 의미를 종합적으로 재구성하여 진술한다. 교육과정과 성취 기준, 평가 기준을 관련지어 보면 다음과 같다.(박귀자, 2003:57-58)



<그림 1> 평가 기준 관련 구분

이 그림에서 '중 수준'이란 '초등학교 6학년' 학생이 충실한 교수·학습 과정을 통해서 성취해야 할 것이라고 기대되는 수준'이라고 할 수 있다. 따라서 중 수준에 도달한 학생은 원칙적으로 초등학교 6학년 학생이라면 누구나 성취하기를 기대하는 필수 내용을 성취했다고 할 수 있다. 그러므로 평가 기준에서 '상 수준'이란 '중 수준을 성취함과 동시에 추가적으로 중 수준보다 심화·발전된 내용을 성취한 수준'이며, '하 수준'은 '중 수준에 해당하는 것을 성취하지 못한 수준'이기 때문에 원칙적으로 초등학교 6학년이면 누구나 성취하기를 기대하는 필수 내용을 제대로 성취하지 못한 수준'이라고 할 수 있다.

2. 말하기 평가의 방법

말하기 평가는 주로 총체적 평가와 분석적 평가 방법을 이용한다. 총체적 평가는 학생의 말하기 수행 능력을 종합적인 관점에서 진술한다. 예를 들어, 말하기를 한 학생에 대하여 '분명하게 대상을 밝히고 그 대상

에 대하여 평가하는 말하기를 하고 있으며, 왜 그러한 평가를 내렸는지 등 평가 대상의 가치가 충분히 드러나도록 그 근거를 확실하게 말하고 있다.'와 같이 종합적인 정보를 제공한다.

분석적 평가는 말하기 능력을 말하기를 구성하는 하위 능력, 즉, '논리성, 타당성, 명료성, 표현성, 청중 의식, 태도' 등과 같은 평가 요소¹⁾로 구분하여 각 평가 요소마다마다 평가 기준을 진술한다. 평가 기준이 분석적인 평가는 학생의 성취 행동에 대하여 장단점을 파악하여 나중에 구체적으로 환류해 줄 수 있고, 각각의 평가 요소별로 기준과 하위 척도가 제공되기 때문에 평가자가 객관적으로 판단할 수 있다.

- 1) 평가 요소는 평가 영역의 구체적인 성질(attribute)을 말하는데 비해, 평가 기준은 평가 요소에 대한 성취 행동의 질적 수준을 판단할 수 있는 근거를 말한다. 읽기를 평가할 때, 추론력을 평가한다면 이는 평가 요소에 해당한다. 이에 비해 평가 기준은 그러한 추론력이 능숙한 경우, 보통인 경우, 미흡한 경우라고 판단할 수 있는 근거와 기준을 말한다. 그러나 많은 평가 상황에서 평가 요소와 평가 기준을 혼동하곤 한다. 다음은 잘못 제시된 평가 기준의 예이다.

평가 기준	배점 및 점수 분포			
1. 주어진 상황에 따른 말하기의 유의점을 고려하였는가?	4점	3점	2점	1점
2. 구체적이고 적절한 내용을 정하였는가?	4점	3점	2점	1점
3. 상황 및 대화의 내용이 분명히 드러났는가?	4점	3점	2점	1점
4. 보조 자료 및 소품의 활용이 적절하였는가?	4점	3점	2점	1점
5. 어법 및 발음, 표정과 태도가 알맞았는가?	4점	3점	2점	1점
〈합계〉	20점			

위의 표에서 평가 기준으로 제시한 것은 실제로는 평가 요소에 해당한다. 이러한 평가 기준으로는 '1. 주어진 상황에 따른 말하기의 유의점을 고려하였는가?'의 요소에 대하여 학생마다 다른 수행을 보일 것인데, 교사가 어느 정도 성취를 보이는 학생에게 4점을 줄 것인지, 또는 3점을 줄 것인지 판단할 수 없다. 교사 개인이 자기가 가르친 학생만 평가한다면 자기 나름의 암묵적 기준으로 판단할 수 있다고 하나 여러 교사가 위와 같은 불명료한 기준을 사용할 경우 평가 결과의 신뢰도를 떨어뜨리며 학생들에게도 어느 정도의 성취를 보여야 4점을 받는지 또는 2점을 받는지 알 수 없다.

Ⅲ. 연구 문제 및 연구 설계

1. 연구 문제

이 연구에서는 실제적인 말하기 평가의 방법으로 활용하는 총체적 평가와 분석적 평가에 따른 평가 방법간 신뢰도를 분석하고자 한다. 이 연구에서 다룰 연구 문제는 다음과 같다.

- 가. 총체적 평가와 분석적 평가에서 부여하는 총점에는 유의한 차이가 있는가?
- 나. 총체적 평가와 분석적 평가에 따른 평가 방법간 신뢰도는 어느 정도인가?

2. 연구 대상 및 기간

연구자가 연구 참여자들이 있는 학교를 직접 방문하여 연구의 내용을 설명하고, 연구에 참여를 원하는 교사들을 대상으로 연구를 실시하였다. 이 연구의 대상은 서울ㄱ초, ㄴ초, ㄷ초, ㄹ초에서 근무하고 있는 교사로, 교직 경력은 만 3년 이상에서 25년까지로 다양하였다. 이를 좀 더 구체적으로 알아보면 다음 <표 1>과 같다.

<표 1> 평가자 실태

학교명 \ 교육경력	3년~5년	6년~15년	16년~25년	계
서울ㄱ초	1	2	1(2)	4(2)
서울ㄴ초	2	3	3(1)	8(1)
서울ㄷ초	1	3	4	8
서울ㄹ초	3	4(1)	4	11(1)
계	7	12(1)	12(1)	31(4)

* ()안은 1차 실험에만 참여한 숫자로, 연구 분석에서 제외.

이 연구는 예비 실험을 2005년 8월 18일부터 20일까지 3일간, 본 실험을 2005년 11월 한 달간 수행하였다.

3. 연구 내용 및 방법

학생들의 '설득하는 말하기' 내용은 2005년 7월 7일 서울초등학교 6학년 국어 수업 시간에 수행평가로 이루어진 것을 캠코더로 녹화한 후 이를 다시 컴퓨터로 편집하여 활용하였다. 6학년 학생들에게는 1주일 전에 설득하는 말하기 주제를 제시해 주었는데, 말하기의 주제는 '초등학생들의 컴퓨터 게임을 규제해야 하는가?', 학생들의 머리를 자율화해야 하는가?, 초등학생들이 학교에 휴대폰을 가지고 다녀도 되는가?, 아파트에서 애완동물을 키워도 되는가?' 등이었다. 발표 시간은 1인당 최대 3분이었다.

이 연구의 예비 실험은 8월 18일부터 20일까지 3일간 약 40여명의 교사들을 대상으로 이루어졌다. 교사 연수 시간을 이용하여 녹화된 화면을 보고 총체적 평가와 분석적 평가를 해 보도록 하였다. 예비 실험에서 교사 평가자들이 지적한 세 가지 문제점은 첫째, 총체적 평가를 한 후에 분석적 평가를 곧바로 하게 되므로, 평가자들이 앞에 평가했던 총체적 평가 점수를 염두에 두고 분석적 평가를 한다는 점, 둘째, 분석적 평가 시에는 실제적으로 두 번 보고 하는 평가이기 때문에 더 정확할 수밖에 없다는 점, 셋째, 너무 많은 학생(16명, 약 28분)을 평가함으로써 평가의 집중도가 떨어진다는 점이었다.

예비 실험의 첫 번째와 두 번째 문제점을 보완하기 위하여 본 실험에서는 연구자가 직접 연구 대상 학교를 각각 2회 방문하였다. 1차 방문에서는 1차에는 연구 내용과 총체적 평가의 평가 요소 등에 대한 설명을 하고 화면을 보면서 총체적 평가만을 실시하였다. 그리고 1차 평가에 대한 기억의 효과가 사라졌을 2주 후에 2차 방문하여 분석적 평가의 평가 요소와 평가 기준에 대한 설명과 함께 분석적 평가를 실시하였다. 예비

실험의 세 번째 문제점을 해결하기 위해 피평가자의 특성이 잘 드러나는 4명만을 선정하여 평가자의 집중도를 높이도록 하였다.

최종적인 4명으로 선정된 학생들의 말하기 내용²⁾은 다음 <표 2>과 같다. 성별간의 인상의 오류를 없애기 위하여 모두 여학생으로 선정하였고, 피평가자의 후광 효과(halo effect)를 없애기 위해 피평가자 학생들의 담임 교사는 평가에서 제외하였다.

<표 2> 피평가자의 말하기 내용

학생	성별	말하기 시간	말하기 주제
A	여	2분 33초	학생들의 머리를 자율화해야 한다.
B	여	38초	초등학생이 휴대폰을 가지도 다녀도 된다.
C	여	1분	아파트에서는 애완견을 키우지 말아야 한다.
D	여	1분 58초	학생들의 머리를 규제하지 말아야 한다.

4. 연구 도구

이 연구를 위하여 도구로 사용한 학생 평가지는 <표 3>, <표 4>와 같다.

<표 3> 총체적 평가지

총체적 평가	말하기 전체에 대하여 평가 요소를 중심으로 평가자가 총체적으로 평가	
평가 요소	발표 준비, 발표 내용 이해, 발표 내용 조직, 표현력, 다른 의견을 가진 사람들에 대한 판단, 다른 사람의 관심을 끄는 능력, 발표 태도	
발표자	발표 특징	점수 (20점 만점)
A		

2) <표 2>에서 보듯 피평가자의 말하기 주제는 각각 다르다. 주제가 다름에 따라 말하기 능력이 달라질 수 있는 변수가 되지만, 이번 실험의 경우에는 미리 주제를 주고 피평가자가 직접 주제를 선정하여 말하기 내용을 준비하였기 때문에 말하기 주제에 대하여서는 변수로 취급하지 않았다.

총체적 평가는 평가 요소를 제시하고 평가 요소를 엄두에 두면서 학생들의 말하기 수행에 대하여 총체적인 점수를 주도록 하였다. 20점 만점에 최고점은 20점, 최저점은 8점으로 점수의 범위를 한정하였다.

<표 4> 분석적 평가지³⁾

분석적 평가	말하기 능력을 하위 요소로 나누어 각각 요소 범주별로 평가		
발표자	평가 요소	점수	총점
A	준비도(발표 준비가 잘 되었는가?)	2, 1, 0	
	이해력(발표 내용(주제)을 잘 이해하고 있는가?)	3, 2, 1, 0	
	조직력(발표 내용을 잘 조직하였는가?)	3, 2, 1, 0	
	표현력(자신의 의견을 제대로 표현하였는가?)	3, 2, 1, 0	
	판단력(다른 의견을 가진 사람들을 고려하고 있는가? 주어진 시간 안에 적절히 발표하는가?)	3, 2, 1, 0	
	의사소통 능력(다른 사람의 관심을 갖도록 하면서 설득력 있게 발표하는가?)	3, 2, 1, 0	
	태도(다른 사람들 앞에서 발표 태도는 바람직한가?)	3, 2, 1, 0	

분석적 평가는 각각의 평가 요소에 따라 점수를 주고, 각 부분 점수를 합한 것을 총점으로 하였다. 최저점이 8점 이상이 되도록 하였으나 총체적 평가와 같이 필수 사항이 아니었다. 왜냐 하면, 최저점을 꼭 한정하게 되면 평가자들이 먼저 총점을 주고 나중에 이 점수에 따라 각 요소

3) 분석적 평가의 평가 요소는 백순근 외(1999:35)의 '구술시험을 위한 평가 기준표(예시)'를 바탕으로 연구자가 재구성하였다.

별 점수를 줄 수도 있었기 때문이다. 분석적 평가는 평가 요소별로 점수를 주는 것이 우선시 되어야 한다.

5. 결과 처리 방법

1차와 2차에 모두 참여한 31명 중에서 1명⁴⁾을 제외하고 30명의 점수는 SPSS Win 10.0 프로그램을 이용하여 전산처리하였다. 이 연구에서 주로 이용한 통계 처리는 총체적 평가와 분석적 평가에 따른 평균과 표준편차, Pearson 상관계수⁵⁾ 분석이었다.

IV. 연구 결과

이 연구에서의 학생 4명에 대한 평가자들의 평가에 따른 연구 결과는 다음과 같다.

1. 학생 A에 대한 분석 결과

학생 A에 대한 기술 통계 결과는 <표 5>와 같다.

-
- 4) 제외된 1명은 학생 평가에서 총체적 평가의 점수를 기준 점수(최하 8점 이상)를 지키지 않은 경우이다.(예: B학생에게 6점을 준 경우) 이 자료는 연구 특성상 신뢰없다고 판단되어 제거한 후 상관계수를 구하였다.
- 5) Pearson 상관계수에 따른 상관계수의 언어적 표현은 다음과 같다.(성태제, 2002: 111)

상관계수 범위	언어적 표현
.00 - .20	매우 낮은 상관이 있다.
.20 - .40	낮은 상관이 있다.
.40 - .60	상관이 있다.
.60 - .80	높은 상관이 있다.
.80 - 1.00	매우 높은 상관이 있다.
1.00	완전한 상관이 있다.

<표 5> 학생 A에 대한 평균, 표준편차

평가 방법	N	Min	Max	Mean	SD
총체적	30	14.00	20.00	17.1667	1.5555
분석적	30	14.00	20.00	16.6333	1.6078

두 평가 방법에 따른 점수의 큰 차이는 없지만 총체적 평가 점수보다는 분석적 평가 점수가 낮았다.(총체적=17.2, 분석적=16.6) 이 결과는 총체적 평가를 실시할 때는 말하기 능력의 수준보다 높은 점수를 주는 경향이 있는 것으로 보이나, 유의확률 .05 수준에서 통계적으로 유의하지는 않았다.($p=.376$)

학생 A에 대한 총체적 평가와 분석적 평가의 상관관계는 <표 6>과 같다.

<표 6> 학생 A에 대한 이변량 상관계수

		총체적	분석적
총체적	Pearson 상관계수	1.000	-.168
	유의확률 (양쪽)	.	.376
	N	30	30
분석적	Pearson 상관계수	-.168	1.000
	유의확률 (양쪽)	.376	.
	N	30	30

총체적 평가와 분석적 평가의 사이에는 -.168의 상관이 있었다. 그러나 이 상관은 .05수준에서 유의하지 않았다.($p=.376$) 그러므로 총체적 평가와 분석적 평가 사이에 신뢰도가 없다고 단정 지을 수 없다. 같은 학생의 말하기 수행에 대하여 평가 방법에 따라 신뢰도가 없는 평가를 한 것이라고 할 수 없다. 하지만 말하기 수행에 대하여 어떤 평가 기준을 가지고 평가해야 하는지에 대한 평가자들의 평가 전문성이 부족함을 어느 정도 보여준다고 할 수 있다.

2. 학생 B에 대한 분석 결과

학생 B에 대한 기술 통계 결과는 <표 7>과 같다.

<표 7> 학생 B에 대한 평균, 표준편차

평가 방법	N	Min	Max	Mean	SD
총체적	30	10.00	17.00	13.5333	2.0634
분석적	30	8.00	16.00	10.5333	2.4877

총체적 평가 점수보다는 분석적 평가 점수가 매우 낮게 나타났다.(총체적 평가 평균=13.5, 분석적 평가 평균=10.5) 이 결과는 총체적 평가를 실시할 때는 말하기 능력의 수준보다 호의적으로 높은 점수를 주는 것을 보여준다. 이 결과는 유의확률 .05 수준에서 통계적으로 유의하였다.($p=.015$)

학생 B에 대한 총체적 평가와 분석적 평가의 상관관계는 <표 8>과 같다.

<표 8> 학생 B에 대한 이변량 상관계수

		총체적	분석적
총체적	Pearson 상관계수	1.000	.440*
	유의확률 (양쪽)	.	.015
	N	30	30
분석적	Pearson 상관계수	.440	1.000
	유의확률 (양쪽)	.015	.
	N	30	30

* 상관계수는 0.05 수준(양쪽)에서 유의함.

총체적 평가와 분석적 평가의 사이에는 유의확률 .05 수준에서 유의한 상관이 있었다.($r=.440$) 이는 총체적 평가와 분석적 평가 사이에 신

뢰도가 있음을 보여준다. 즉, 학생 B의 말하기 수행을 평가하는 기준 사이에는 평가 방법간에 신뢰도가 있음을 나타내는 것이다. 학생 B의 경우 네 명의 학생 중에서 두 평가 방법에서 모두 가장 낮은 평균 점수를 얻었다. 말하기를 못하는 학생에 대해서는 평가 방법에 따른 평가의 신뢰도가 있음을 보여주는 결과이다. 즉, 학생 B는 유의성이 있으므로 정적(positive)인 상관관계가 있다.

3. 학생 C에 대한 분석 결과

학생 C에 대한 기술 통계 결과는 <표 9>과 같다.

<표 9> 학생 C에 대한 평균, 표준편차

평가 방법	N	Min	Max	Mean	SD
총체적	30	8.00	19.00	14.4667	2.5695
분석적	30	8.00	18.00	12.7667	2.9906

총체적 평가 점수보다는 분석적 평가 점수가 낮게 나타났다.(총체적 평가 평균 $\approx$ 14.5, 분석적 평가 평균 $\approx$ 12.8) 이는 총체적 평가를 할 때는 분석적 평가를 할 때보다 좀 더 높은 점수를 경향을 보여주고 있으나, 이 결과는 유의확률 .05 수준에서 통계적으로 유의하지 않았다.($p=.260$)

학생 C에 대한 총체적 평가와 분석적 평가의 상관관계는 <표 10>과 같다.

<표 10> 학생 C에 대한 이변량 상관계수

총체적		총체적	분석적
	Pearson 상관계수	1.000	.212
	유의확률 (양측)	.	.260
	N	30	30

분석적	Pearson 상관계수	.212	1.000
	유의확률 (양쪽)	.260	.
	N	30	30

총체적 평가와 분석적 평가의 사이에는 .212의 상관이 있었다. 그러나 이 상관은 유의확률 .05수준에서 유의하지 않았다.($p=.260$) 그러므로 총체적 평가와 분석적 평가 사이에 신뢰도가 없다고 단정 지을 수 없다. 같은 학생의 말하기 수행에 대하여 평가 방법에 따라 신뢰도가 없는 평가를 한 것이라고 할 수 없다. 학생 C의 경우 네 명의 학생 중에서 두 평가 방법에서 모두 두 번째로 낮은 평균 점수를 얻었다. 그러나 학생 B의 경우는 유의성이 있으므로 정적(positive)인 상관관계가 있다고 할 수 있으나, 학생 C의 경우는 -(음) 상관관계가 나타났지만, 유의성이 없기 때문에 부적(negative)인 상관관계라고 할 수 없다.

4. 학생 D에 대한 분석 결과

학생 D에 대한 기술 통계 결과는 <표 11>과 같다.

<표 11> 학생 D에 대한 평균, 표준편차

평가 방법	N	Min	Max	Mean	SD
총체적	30	17.00	20.00	18.6333	.9643
분석적	30	17.00	20.00	19.2000	.8867

분석적 평가 점수가 총체적 평가 점수보다 조금 높게 나타났다.(총체적 평가 평균 $\approx$ 18.6, 분석적 평가 평균 $\approx$ 19.2) 이는 앞의 학생 A, B, C와는 다르게 분석적 평가를 할 때, 총체적 평가를 할 때보다 좀 더 높은 점수를 주고 있음을 보여준다. 그러나 이 결과는 유의확률 .05 수준에서 통계적으로 유의하지 않았다.($p=.703$)

학생 D에 대한 총체적 평가와 분석적 평가의 상관관계는 <표 12>와 같다.

<표 12> 학생 D에 대한 이변량 상관계수

		총체적	분석적
총체적	Pearson 상관계수	1.000	-.073
	유의확률 (양쪽)	.	.703
	N	30	30
분석적	Pearson 상관계수	-.073	1.000
	유의확률 (양쪽)	.703	.
	N	30	30

총체적 평가와 분석적 평가의 사이에는 -.073의 상관관계가 있었다. 그러나 이는 유의확률 .05 수준에서 유의하지 않았다. 그러므로 총체적 평가와 분석적 평가 사이에 신뢰도의 유무는 알 수 없다. 학생 D의 말하기 수행을 평가하는 기준 사이에도 평가 방법간에 -(음) 상관관계가 나타났다. 학생 D의 경우, 학생 A와 마찬가지로 네 명의 학생 중에서 두 평가 방법에서 모두 높은 평균 점수를 얻은 학생이었다. 평균적으로 높은 점수를 얻은 학생들의 평가에서 평가 방법간에 -(음) 상관관계가 나타났지만, 모두 유의성이 없기 때문에 부적(negative)인 상관관계라고 할 수 없다.

V. 결론 및 제언

말하기 평가는 실제적이어야 한다. 왜냐 하면 실제적 평가가 타당하기 때문이다. 그러나 말하기 평가 방법이 타당도는 알맞지만 신뢰도가 낮다면 알맞은 말하기 평가로 활용되기 어렵다. 말하기의 실제적 평가가

올바르게 정착되기 위해서는 평가의 타당도와 함께 평가의 신뢰도도 확보되어야 한다.

이 연구는 초등학교 6학년 학생들의 설득하는 말하기 수행에 대한 총체적 평가와 분석적 평가 방법에 따른 평가 방법간의 신뢰도를 분석하였다. 연구 결론은 다음과 같다.

첫째, 분석적 평가에 비하여 총체적 평가의 점수를 높게 주는 경향이 있다. 학생 D의 경우를 제외하고는 총체적 평가의 점수가 높았다. 하지만 학생 B의 경우에만 통계적으로 유의하였다. ($p=.015$) 이는 총체적 평가에서 분석적 평가보다 높은 점수를 준다는 일반적인 생각을 부분적으로 증명해 주는 것이라 할 수 있다.

둘째, 말하기 수행에 대한 총체적 평가와 분석적 평가의 평가 방법간의 신뢰도는 학생 B의 경우에만 유의한 상관(정적인 상관)이 존재하였다. 다른 A, C, D 세 학생에 대해서는 유의확률 .05 수준에서 유의성이 없으므로 정적인 상관관계 또는 부적인 상관관계가 있다고 할 수 없다. 이는 평가 방법에 따른 신뢰도를 높이기 위한 노력이 뒤따라야 함을 보여준다. 이와 같이 평가 방법에 따른 평가의 신뢰도가 4명 중 1명의 경우에만 정적인 상관관계를 보인다면 대규모 평가를 대신하여 말하기를 비롯한 국어과 평가를 실제적 평가로 실시할 때의 큰 문제점이라 할 수 있다. 실제적 말하기 평가의 신뢰성이 없으므로 고부담 평가(high stakes assessment) 상황에서 활용하기는 더더욱 어렵다고 할 수 있다.

유의한 상관이 적다는 것은 총체적 평가든 분석적 평가든 평가를 실시하기 전에 반드시 평가자들의 연수나 협의와 훈련이 필요하다는 것을 보여주는 것이다. 이는 이인제 외(2004)의 교사의 학생 평가 전문성 신장 모형에서 제시한 학교, 시·도 교육청, 국가 수준의 평가 전문성 신장을 위한 내용과 방법이라는 큰 틀과 부합되는 것이라 할 수 있다.

셋째, 특히, 말하기 평균 점수가 높은 학생 A, D의 경우에는 -(음) 상관관계(유의하지 않지만)를 보여, 말하기 평균 점수가 낮은 B, C의 경우와는 대비되었다. 이는 말하기를 못하는 학생에 대한 평가 방법에는 부분적으로 신뢰도가 있음을 보여주고 있으나 말하기를 잘하는 학생에 대

해서는 평가 방법에 따라 신뢰도가 없음을 보여주었다. 이는 말하기 평가 요소와 평가 기준이 무엇인지에 대한 사전 협의가 필요하며, 각각의 평가 요소에 따른 평가 기준이 무엇인지 세밀하면서 알기 쉽게 제시해야 할 필요가 있음을 보여주는 것이다. 특히 말하기를 잘하는 학생에 대한 평가 기준 설정 등에 더욱 깊은 연구가 필요함을 보여주는 것이다.

연구 결론을 바탕으로 한 제언은 다음과 같다.

첫째, 교사들의 평가자 연수와 훈련이 필요하다. 말하기 수행에 대한 평가 방법에 따른 평가 결과의 신뢰도가 높도록(일반적으로 $r = .8$ 이상) 평가자 연수와 훈련을 실시해야 한다. 평가자내 신뢰도(intra-rater reliability)를 확보하기 위한 연수와 훈련이 있어야 한다. 그리고 평가 방법간, 그리고 평가자간 신뢰도(inter-rater reliability)를 분석하여 평가자 훈련의 효과를 검증해야 한다.

둘째, 실제적 말하기 수행 능력을 제대로 평가할 수 있는 평가 기준이 설정되어야 한다. 이 연구에서 평가의 신뢰도가 낮게 나온 원인 중의 하나는 평가자들이 서로 다른 평가 기준으로 평가하기 때문이다. 그러므로 델파이 기법 등을 이용하여 평가 요소와 이에 따른 평가 기준(채점 기준)을 상세하게 하여 평가자의 주관성이 최대한 배제되도록 해야 한다.

셋째, 학생들이 평가자가 되었을 때의 평가자 신뢰도를 검증하는 연구가 이어져야 한다. 특히 초등학생들의 경우 학생 상호 평가에서의 신뢰도는 큰 문제가 될 수 있다. 평가 방법적인 측면에서 자기 평가나 학생 상호평가를 활용하는 것이 교사 평가의 주관성이나 비신뢰성을 보완할 수 있다고 하였지만 학생들의 평가 신뢰도가 낮다면 학생들의 자기 평가나 상호 평가를 평가의 방법으로 활용하기 어렵기 때문이다.*

* 본 논문은 2006. 10. 19. 투고되었으며, 2006. 11. 17. 심사가 시작되어 2006. 11. 27. 심사가 종료되었음.

■ 참고문헌

- 김선배·이도영·장일범(1998), 초등국어과 말하기 영역의 수행평가 모형 개발 및 현장 적용 연구, 대학과 협력 연구 논문 제8집, 강원도 교육과학연구원.
- 노명완(1992), 말하기 및 듣기의 본질과 평가의 방향, <제3회 국어교육연구 학술발표대회 자료집>, 한국국어교육연구회.
- 류성기(1996), 국어 말하기 교육의 효율적인 평가 방안, <교육 한글> 9, 71-122, 한글학회.
- 박귀자(2003), 설득하는 말하기에 관한 연구, 부산교대 석사학위논문.
- 박영목(1992), 중·고등학교에서의 말하기·듣기 평가, <제3회 국어교육연구 학술발표대회 자료집>, 한국국어교육연구회.
- 박영목(1999), 작문 능력 평가 방법과 절차, <국어교육> 99, 1-29, 한국국어교육연구회.
- 백순근·최미숙(1999), 고등학교 국어과 수행평가의 이론과 실제, 한국교육과정평가원.
- 성태제(2002), 현대 기초통계학의 이해와 적용(개정판), 서울: 교육과학사.
- 신동일(2006), 한국의 영어평가학2: 말하기 시험편, 서울: 한국문화사.
- 신현재·원진숙·천경록·김도남·임천택·서현석(2003), 생태학적 관점에 의한 국어과 평가 방안 연구, 교과교육공동연구지원사업 B-00036 연구보고서.
- 원진숙(1997), 말하기·듣기 영역의 평가, <97 국어교육연구소 학술발표회 자료집>, 서울대학교 국어교육연구소.
- 이도영(2001), 국어과 교육 평가 모형, <국어교육학연구> 13, 297-320, 국어교육학회.
- 이도영(2005), 말하기·듣기 수업 평가 기준 마련을 위한 시론: 교수·학습 방법을 중심으로, <한국초등국어교육> 27, 21-43, 한국초등국어교육학회.
- 이인제·정구향·천경록·이도영·김창원·이성영·심영택·김신영·채선희(1997), 창의력 신장을 돕는 중학교 국어과 학습 평가 방법 연구, 연구보고 CR 97-10, 한국교육개발원.
- 이인제·이범홍·박정·진재관·김옥남·서수현·김신영(2004), 교사의 학생 평가 전문성 신장 모형과 기준, 연구보고 RRE 2004-5-2, 한국교육과정평가원.
- 이창덕(1999), 대화의 메타 분석을 활용한 말하기 평가 방법 탐색, <99 국어교육연구소 정기 학술대회 발표집>, 서울대학교 국어교육연구소.
- 조재운(2004), 프리젠테이션 교육의 문제점과 그 개선 방향, <화법연구> 7, 93-116, 한국화법학회.
- 조재운(2005), 말하기 평가의 실제, <초등국어교육> 36, 27-68, 서울초등국어교과교육연구회.
- 조재운(2006), 국어과 말하기·듣기 평가 실태 조사 연구, <화법연구> 9, 115-141, 한국화법학회.
- Chaney, A. L., T. L. Burk(1998), *Teaching Oral Communication in Grades K-8*, Allyn & Bacon.

- Cooper, J. D., N. D. Kiger(2001), *Literacy Assessment: Helping Teachers Plan Instruction*. Houghton Mifflin Company.
- Corden, R.(2000). *Literacy and learning through talk: Strategies for the primary classroom*. Buckingham: Open University Press.
- Darling-Hammond, L., J. Ancess. & B. Falk.(1995), *Authentic Assessment in Action: Studies of Schools and Students at Work*. Teachers College Press.
- Fehring, H.(Ed.)(2003). *Literacy Assessment*. IRA.
- Jett-Simpson, M., L. Leslie.(1997), *Authentic Literacy Assessment: An Ecological Approach*. Wisconsin State Reading Association.
- McMillan, J. H.(2004), *Classroom assessment: Principles and Practice for Effective Instruction*.(3rd) Pearson.
- Valencia, S. W., E. H. Hiebert, P. P. Afflerbach.(Eds)(1996),. *Authentic Reading Assessment: Practices and Possibilities*. IRA.

<초록>

말하기 평가 방법간 신뢰도 분석 연구

— 초등학교의 설득하는 말하기를 중심으로 —

조재윤

이 연구의 목적은 초등학교 6학년 학생들이 수행한 설득하는 말하기에 대한 교사 평가자들의 총체적 평가와 분석적 평가 방법간 신뢰도를 분석하는 것이다.

총체적 평가는 말하기 능력을 종합적인 관점에서 진술하는 것인 반면, 분석적 평가는 말하기를 구성하는 하위 능력, 즉, '논리성, 타당성, 명료성, 표현성, 태도' 등의 평가 요소로 구분하여 각 평가 요소마다 평가 기준을 진술한다.

연구 결과, 평가자들이 분석적 평가보다 총체적 평가의 점수를 높게 주는 경향이 있으나 학생 B의 경우에만 통계적으로 유의하였다. ($p=.015$) 말하기 수행에 대한 총체적 평가와 분석적 평가의 평가 방법간의 신뢰도는 학생 B의 경우에만 유의한 상관이었다.

말하기 평가 방법간 신뢰도를 높이기 위해서는 평가 방법에 상관없이 평가 실시 전에 평가자들의 연수가 필요하고, 실제적인 말하기 평가 요소와 구체적이면서 적용하기 쉬운 평가 기준 설정에 대한 연구가 있어야 한다.

【핵심어】 말하기 평가, 말하기 평가의 신뢰도, 총체적 평가, 분석적 평가, 평가 기준, 설득하는 말하기

<Abstract>

An Analytical Study on Reliability of Speaking Assessment Methods

Cho, Jae-yoon

The purpose of this study is to analyze reliability of holistic and analytical assessments teachers as raters implemented on sixth-graders' persuasive speaking.

Raters tended to give higher marks in holistic assessment than in analytical assessment. All the students but D got higher marks in holistic assessment, which was statistically significant for B alone ($p=.015$). Only B showed significant correlation in reliability between holistic assessment and analytical assessment of speaking.

To improve reliability of speaking assessment methods, it is essential to train raters. Also necessary is a research in authentic speaking assessment elements as the foundation for speaking assessment and in setting assessment standards which is easy to apply.

[Key words] speaking assessment, reliability of speaking assessment, holistic assessment, analytical assessment, assessment standards, persuasive speaking.