중등학교 영어과 평가 개혁의 방향

(Directions of Test Reform in the Secondary English Education)

강성우*

요약

본 연구는 평가의 결과가 중등학교 영어교육이 교육부에서 정한 영어과 교육과정 목표를 제대로 달성하고 있는지를 판단하는 근거가 될 수 있는가의 관점에서 현재 널리 시행되고 있는 다지선다형 평가방식의 한계점에 대해 살펴보고 이에 대한 대안을 제시하고자 한다. 교과과정 목표의 달성 여부를 판단하기 위해서는 평가로부터 학생들의 능력을 직접적으로 판단할 수 있어야 한다. 그러나 현재 시행되고 있는 다지선다형 평가는 학생들의 능력의 순위에 대한 정보를 제공하지 학생들의 영어 능력에 대한 정보를 제시하지 않는다. 이에 대한 대안으로 학급 단위에서는 수행평가를 이용하여 학습자의 능력을 구체적으로 기술할 것과, 중등학생 전체를 파악하기 위해서는 문항 반응 이론을 이용한 다지선다형 평가를 이용할 것을 제안한다.

1. 서론

현대 사회에서 제2 외국어 특히 영어를 잘 사용하는 능력의 중요성은 더 이상 언급할 필요조차 없는 누구나 느끼는 사실이 되어버렸다. 이러한 영어 사용능력의 필요성에 발 맞추어, 일선 중고등학교의 교사와 학생과 학부형들의 관심도 이전에 비하여 실제로 언어를 사용할

* 청주교육대학교 영어교육과

수 있는 능력을 배우고 가르치는 쪽으로 많이 기울어졌다. 이런 변화는 교육 정책상에서도 일어났다. 과거의 영어 교육에서 교과 과정상의 목표가 듣기, 쓰기, 말하기, 읽기 등의 네 기능의 균형 있는 발달로 설정이 되어 있었지만 실제의 교실 현장에서는 듣기와 읽기 능력만을 측정해왔던 고등학교와 대학교 입학시험에서 좋은 점수를 받기 위한 교육을 하였다고 할 수 있다. 2001년부터 단계적으로 시행되는 제7차 교육과정에서는 교실 현장에서 보다 실질적인 학습내용과 방법상의 변화를 이끌어 내기 위하여, 단계별 학습을 도입하였다. 이에 따르면 학생들은 상급단계로 진급하기 위해서는 각 단계별로 일정한 능력을 갖추어야 한다는 것이다. 이러한 단계별 학습 정책이 성공적으로 시행되기 위해서 먼저 해결되어야 하는 것이 어떻게 학생들의 능력을 평가하느냐 하는 문제이다. 만약 단순히 이전에 시행되던 식의 듣기와 읽기 위주의 평가가 행하여지면, 학생들은 읽기와 듣기 이외의 기능에 대하여 학습할 동기를 가지지 못할 것이다. 이 연구에서는 지금까지 시행되어 왔던 다지선다형의 문제점을 지적하고, 제7차 교육과정을 성공적으로 시행하기 위해 중등교육에서의 영어평가를 어떻게 바꾸어야 할 지에 대하여 논하여 보고자 한다.

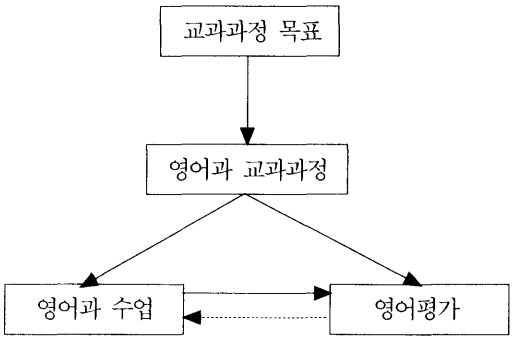
2. 평가의 기능

현재 행하여지는 평가의 문제점들을 언급하기 이전에, 교육제도 내에서 언어 평가가 수행하는 기능에 대하여 생각해 볼 필요가 있다. 평가의 기능은 학생, 교사, 교육제도 3가지 측면에서 살펴볼 수 있다. 첫째, 평가는 학생들의 능력과 학업 성취 정도를 측정한다. 둘째, 평가는 교사의 수업에 대한 평가기능을 수행한다. 즉, 교사의 수업이 얼마나 효율적이었나에 대한 평가의 수단이 될 수 있다. 셋째, 평가에서 얻어

지는 정보는 교과 과정 목표의 달성여부를 판단하는 근거가 될 수 있다. 본 연구에서는 위의 세 번째 기능에 초점을 두어, 현재 행하여지는 평가로는 학생들이 교과과정에서 서술하는 목표의 달성 여부를 판단할 수 없음을 보여주려 한다. 이러한 기능은 제7차 교육과정에서부터 도입되는 단계학습의 성공여부와 직결되는 문제이다. 평가의 이러한 기능을 보다 잘 이해하기 위하여 교육제도 내에서의 평가의 위치를 이해할 필요가 있다.

학교에서 행하여지는 다른 모든 평가와 마찬가지로, 영어 평가는 교과과정과 교과 수업에 연관시켜 그 위치를 생각할 수 있다. [그림 1]은 교육제도 내에서의 영어 평가의 위치를 보여준다. 교과 과정은 국가에서 국가적 사회적 필요성에 의해 정한 교과과정 목표에 맞추어 정하여지며 학교에서의 영어과 수업은 교과 과정에 준해 이루어진다. 학교에서 행하여지는 영어 평가는 두 가지로 나누어 볼 수 있다. 하나는 일선 교사가 만드는 학생들이 수업시간에 얼마나 학습하였는가를 측정하는 평가이고, 다른 하나는 학생 전체를 대상으로 만들어져, 학생들이 교과과정에서 정한 학습목표에 얼마나 도달하였는가 하는 평가이다. 이때 학과 수업이 교과과정 목표를 잘 반영하면, 교과 과정에 준한 평가나 수업에 준한 평가는 모두 같은 능력을 측정한다고 할 수 있다. 그러나 교과 수업이 교과과정 목표를 제대로 반영하지 못할 경우, 두 평가는 측정하는 능력에 있어서 차이를 보이게 된다. 즉 교사에 의해 준비된 성취도 평가로서는 교과과정에서 목표하는 능력을 측정할 수 없으며, 교과과정에 준한 평가는 학생들이 배우지 않은 능력을 측정하게 된다.

다음과 같이 교과 수업이 교과과정을 제대로 반영하지 못한 상황에서 교과과정에 준한 평가는 교과 수업에 긍정적 또는 부정적으로 막대한 영향을 끼치고 변화시킬 수 있다. 이러한 평가의 기능을 파급효과라 한다(McNamara, 2000). [그림 1]의 '영어 평가'로부터 '영어과 수업'으로 향하는 점선 화살표는 평가의 이러한 기능을 나타낸다. 특



[그림 1] 교육제도 내에서의 영어 평가의 위치

히 우리 나라에서처럼 평가 결과에 따라서 입학이나 진로 등의 많은 결정이 이루어지는 상황에서는 평가가 어떻게 만들어지는가에 따라 교과 수업 내용이 크게 달라진다. 본 연구에서 목표하는 것이 바로 올바른 평가의 방향을 설정함으로써 학생과 교사의 관심과 수업 방향이 진정한 영어 능력을 향상시키는 쪽으로 향하도록 하는 것이다. 이를 위해 먼저 현재 행하여지는 평가로서는 학생들이 교과과정에서 정한 목표를 달성하였는지를 판단하기에 충분한 정보를 얻을 수 없는 이유에 대해 살펴보겠다.

3. 평가 결과와 교육과정 목표의 불일치

이전의 교육과정과 마찬가지로 제7차 교육과정도 영어과 교육과정의 목표를 "일상생활에 필요한 영어를 이해하고 사용할 수 있는 기본적인 의사소통 능력을 기른다"로 하고 있으며 듣기, 읽기, 쓰기, 말하기 등의 네 기능의 개별적 능력과 함께 네 기능의 통합적 사용능력을 함양시킬 것을 그 세부 내용으로 하고 있다(교육부, 1998). 초등학교

에서부터 고등학교 1학년까지 적용되는 국민 공통 기본 교육과정의 영어과 교육 내용의 편성과 운영을 보면, 초등학교 3학년에서 6학년까지의 학생을 대상으로 하는 심화보충형 수준별 교육과 중학교 1학년부터 고등학교 1학년 학생들에게 적용되는 단계별 교육과정을 운영하도록 되어 있다. 각 단계별로는 적절한 성취기준이 마련되어 있으며 학생들이 상위 단계의 수업을 받기 이전에 하위 단계의 성취 기준에 맞는 능력을 갖출 것을 명시하고 있다.

이렇게 단계학습 개념을 도입한 제7차 교육과정의 성공적 시행을 위해서는 무엇보다도 중요한 것이 교육과정 목표에 적절한 평가—각 단계별 성취도를 판단할 수 있는 평가—를 만드는 것이다. 적절한 평가가 어떠한 것인가를 알아보기 위해 현재 시행되는 평가들이 적합하지 않은 이유를 살펴볼 필요가 있다.

현재 우리 나라의 중고등학생들에게 시행되는 시험의 형태는 대부분이 다지선다형이다. 학생들은 각 문항에서 주어진 몇 가지 답안에서 정답과 오답을 가려내도록 요구된다. 이러한 평가들은 만들어지는 과정별로 크게 두 가지로 나눌 수 있다. 하나는 교사가 자신의 수업에 준해서 만드는 평가로 학교에서 시행되는 평가의 대부분이 이에 해당한다고 할 수 있으며, 학생들이 수업내용을 얼마나 학습하였는가를 측정한다. 다른 하나는 전체 학생을 대상으로 하는 평가로서 사설 평가기관이나 국립 평가기관에 의해 만들어지는 표준화 평가이다. 이러한 평가중 대표적인 것이 예전의 대학입학 학력고사나 오늘날의 수능평가이다. 일선 학교에서 이용되는 공인된 교과서가 여러 종류이므로, 전체 학생 대상 표준화 평가는 교과서 편찬시 기준이 되는 교육과정에 준하여 만들어져 학생들이 교육과정에서 정한 성취목표를 얼마나 달성하였는가를 측정한다.

이들 평가들의 문제점을 크게 두 측면에서 생각하여 볼 수 있다. 첫째는 위에서 언급한 평가의 세 번째 기능 측면으로, 이들 평가로서는

교육과정 목표가 얼마나 달성하였는지 판단할 수가 없다. 영어과 교과과정 목표에는 듣기, 말하기, 쓰기, 읽기 등의 네 언어기능의 균형 있는 발달을 이루도록 서술되어 있으나, 현 평가의 형식인 다지선다형으로서는 읽기와 듣기 등의 언어의 수용적 기능만을 측정할 뿐이고, 말하기와 쓰기 등의 산출적 기능을 측정할 수 없다. 즉, 영어과 교과과정 목표의 달성여부에 대한 충분한 정보를 제공하지 못한다. 평가가 수용적 언어 기능만을 측정하는 것은 수업내용에도 영향을 주어왔다. 우리의 교육제도에서 대학입시가 차지하는 비중을 고려해볼 때, 학생들이 시험에서 높은 점수를 받도록 일선학교 교실의 수업이 이루어지는 것은 너무도 당연하다. 따라서 쓰기 교육이나 말하기 교육은 중고등학교 수업에서 소홀히 되어왔다.

다지선다형 평가는 종종 평가하고자 하는 목표 지식이 작은 세부적 지식들의 합이라는 가정하에 만들어진다. 그래서 학생들의 종합적인 언어사용 능력보다는--어떤 글로부터 추론을 하거나 주제를 찾거나 하는 능력--문법사항이나 단어 등의 단편적인 언어지식을 묻는 문제들이 많이 만들어져 왔다. 다음은 이러한 읽기기능을 측정하는 문제들의 일례이다(한국교육과정평가원(출판예정)에서 재인용).

It is not necessary to kill whales. There are other products to use instead of whale oil and meat. If the killing (continue), the wonderful animal will disappear from earch. In the longrun, there will be no whales in world.

( ) 안의 continue의 알맞은 형태를 고르시오.

1) continued 2) continues 3) to be continued 4) is continuing

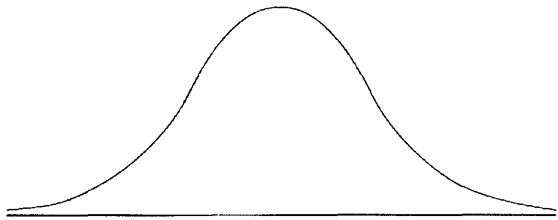
읽기라 하면 어떤 글을 통해 정보를 얻고 지식을 증진시키는 것이

라 할 수 있다. 하지만 위의 예와 같은 문제는 학생들이 실제 세계에서 글을 읽고 정보를 얻는 것과 같은 기능을 수행하도록 하는 것이 아니라, 내용과는 별반 관계없이 조건문에 적합한 동사의 형태만을 고르도록 한다. 이러한 평가 문항들은 학생들의 관심을 영어의 사용보다는 영어의 용법에만 관심을 가지게 되어 학생들이 의사소통 능력을 기르는데 도움이 되지 못 하였다.

두 번째 문제점은 현재의 다지선다형 평가의 배경이론상의 한계로 인해 평가의 이용도 면에서 많은 제약을 받는다. 지금까지의 교사의 수업에 준한 평가나 교육과정에 준한 표준화 평가들은 고전 평가 이론(classical test theory)에 근거하여 만들어 졌다. 고전 평가 이론에 의한 평가는 4가지의 특징을 가진다. 첫째, 고전 평가 이론은 평가의 신뢰도와 타당도를 매우 중요히 여겨 이에 대한 많은 연구를 이루었다. 신뢰도는 평가 결과의 일관성이다. 즉, A라는 능력을 가진 수험생의 평가 결과는 시험을 오늘 치르건 아니면 한달 후에 치르건 능력의 변화가 없는 이상에는 항상 같은 점수가 결과로 나와야 그 평가가 신뢰도가 있다고 할 수 있다. 타당도는 평가가 측정하고자 목표한 것을 얼마나 잘 측정하는가의 정도를 나타낸다. 즉, 영어시험이라면, 단순히 영어 능력만을 측정해야지, 영어로 된 복잡한 추리 문제가 학생들에게 주어진다면 그 시험은 영어 능력뿐만 아니라 추리 능력까지도 측정하는 것이 된다. 한편, 영어 능력을 측정하는 시험이 듣기와 읽기 능력만 측정한다면 이도 또한 타당한 영어 능력 평가라 할 수 없는 것이다. 신뢰도와 타당도는 고전 평가 이론에서만 중시한 것이 아니라 모든 평가가 갖추어야 하는 특성이지만 특히 고전 평가 이론 내에서 많은 발전을 하여 다지선다형 문항에 대한 수험자의 반응을 기초로 두 개념을 수치적으로 계산할 수 있게 되었다.

두 번째 특징은 고전 평가 이론에서 평가의 목적은 학생들의 능력에 대한 정보를 얻는 것이 아니라 학생들의 능력을 구분하는 것이다.

즉 학생들이 무엇을 할 수 있고 무엇을 할 수 없는 지를 말하는 것이 아니라 성적이 좋은 학생과 성적이 나쁜 학생을 구분하는 것이다. 이러한 특징은 고전 평가 이론의 세 번째 특징인 평가 결과의 정상분포와 관련 있다. 고전 평가 이론에 따르면 제대로 만들어진 평가의 결과는 정상분포를 이루어야 한다. [그림 2]는 정상분포 곡선이다. 가로축은 학생들의 점수를 나타내며 곡선의 높이는 각 점수대 학생 수의 상대적 분포를 나타낸다. 중간 성적의 학생들의 분포가 많으며, 성적이 낮아질수록 그리고 높아질수록 학생들의 분포가 상대적으로 적어지게 된다. 시험 결과가 정상분포로 나타나는 한, 시험을 치른 전체 학생들의 능력 수준은 문제가 되지 않는다. 개개 학생에 대한 평가는 그 학생의 성적이 전체 수험자중 몇 등인가에 의해서 이루어진다. 즉, 평가를 통해서 학생들이 영어 원어민과 간단한 의사소통을 할 수 있는지를 판단하는 것이 아니라 학생의 성적이 상위 10%에 속하는지 20%에 속하는지를 판단하는 것이다.



[그림 2] 정상분포 곡선

네 번째로, 좋은 평가를 만들기 위해서는 좋은 문항들을 만들어야 한다고 여겨졌다. 이러한 생각은 문항 분석의 개념을 낳았고 각 문항의 평이도와 변별도라는 개념을 발전시켰다. 평이도는 그 문항이 쉬운 정도를 나타내는 것으로 전체 수험자중 해당 문항에 올바르게 답을

한 수험자의 비율로 나타낸다. 평이도가 0.8이면 80%의 수험자가 올바른 답을 한 것이다. 변별도는 한 문항이 잘하는 학생과 못하는 학생을 구분하는 정도를 나타낸다. 즉 높은 시험점수를 받은 학생 중 상당수가 어떤 문항에 올바르게 답을 하지 못하였는데, 낮은 시험점수를 받은 학생들 대부분이 그 문항에 올바르게 답을 하였다면 그 문항의 변별도는 대단히 낮은 것이다. 시험 결과가 정상분포로 나오도록 하기 위해서는 문항들의 평이도를 적절히 조정하여야 하며, 높은 변별도를 지닌 문항들로 이루어진 평가를 구성하여야 한다. 하지만 정상 분포의 지나친 강조는 평가 문항의 내용보다는 문항에 대한 반응의 수치적 특성을 평가 문항 선택의 일차적 조건으로 여기는 오류를 낳았다 (Davidson, 1993; Peirce, 1992).

4. 고전 평가 이론의 한계

지금까지 살펴본 고전 평가 이론은 신뢰도, 타당도, 변별도, 평이도 등 많은 유용한 개념들을 발전시켜, 보다 객관성 있는 평가 구성을 위한 많은 발전을 이루었지만, 객관성에 대한 지나친 강조와 수험자들의 능력을 파악하는 것이 아니라 석차를 파악하려하였기 때문에, 교과 과정 정책의 수행과 평가에 필요한 정보를 제공하지 못한다. 우리가 얻고자 하는 정보의 관점에서 고전 평가 이론의 한계를 살펴보겠다.

첫째, 가장 중요한 한계점은 평가를 통하여 얻을 수 있는 정보가 수험자가 어느 정도의 능력을 갖추었는지가 아니라 수험자가 전체 수험자에 비교해서 상대적으로 몇 등에 해당하는 정보만을 얻을 수 있다. 둘째, 평가 결과를 전체 수험자와 독립적으로 해석할 수 없다. 예를 들어, 같은 능력을 가진 A와 B 두 수험자 중 A는 시험을 3월에 치렀고, B는 5월에 치렀다고 생각하여 보자. 두 사람의 능력이 같다면 평

가 결과가 같게 나타나야만 제대로 된 평가라 할 수 있다. 하지만 고전 평가 이론에서는 수험자의 능력이 상대적으로 평가되기 때문에 3월에는 A와 B보다 성적이 좋은 학생들이 많이 시험을 치렀지만 5월에는 A와 B보다 성적이 좋지 않은 학생들이 많이 시험을 치렀다면, A와 B에 대한 평가는 달라지게 된다. 셋째, 앞의 두 번째 한계점은 평가 결과의 이용을 단발성에 그치게 한다. 즉 평가결과가 수험자들이 어떤 사람들이었는지에 따라 다르게 나오기 때문에, 앞의 경우에서 3월의 시험과 5월의 시험 결과를 비교할 수 없다.

이러한 문제점들 때문에 고전 평가이론에 바탕을 둔 다지선다형 평가는 학습자의 능력에 대한 신뢰할 만한 지표를 제공하지 못한다. 즉 교육과정 목표의 달성여부를 판단하기 위해 필요한 정보(학생들이 어느 정도의 성취를 이루었나에 대한 정보)를 제공하지 못한다. 평가가 교육과정 정책의 수행과 검증에 이용되기 위해서는 평가 결과가 수험자의 능력에 대해 직접적으로 지시하여야 하며, 같은 능력을 측정하는 평가들간에 결과를 비교할 수 있어야 한다.

5. 제안

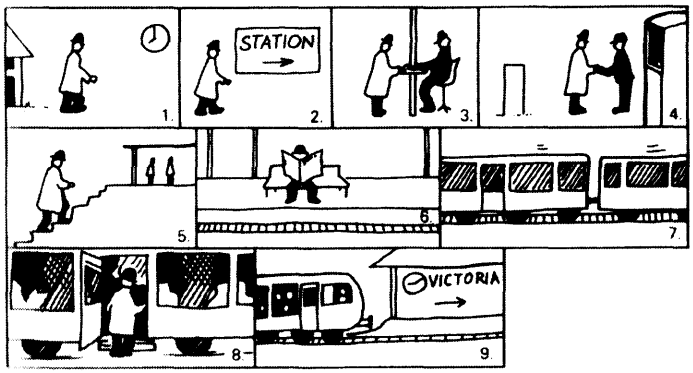
교육과정 정책은 전체학생을 대상으로 하는 것이기 때문에, 교육과정 정책에 대한 검증을 하기 위해서는 전체 학생을 대상으로 학업정도를 평가하여야 할 필요가 있다. 따라서, 우리가 필요한 평가는 앞에서 언급한 두 조건들뿐만 아니라, 학생들 전체를 대상으로 하는 대규모의 시행이 가능하여야 한다. 하지만 이 조건들을 모두 만족시키는 평가를 만들어 시행하는 것은 대단히 어렵다. 학습자의 언어 능력의 정도를 정확하게 나타내기 위해서는 듣기, 말하기, 읽기, 쓰기 등의 언어 기능의 종합적 사용능력을 측정하여야 하며, 이에는 영어사용의 직

접적 관찰이 필요하다. 하지만 이러한 평가를 대규모로 시행하기에는 너무나 많은 인력과 재원이 요구된다. 대안으로서 본 연구에서는 평가를 교실에서의 평가와 전체학생 대상의 평가로 이원화하여 새로운 평가를 구성할 것을 제안한다.

5.1. 교실에서의 준거 참조 평가(criterion-referenced testing)

이전의 다지선다형 시험은 평가 결과를 정상 분포를 이루도록 평가를 만들려고 하였던 규준 참조 평가(norm-referenced testing)이었다. 이에 반해 준거 참조 수행평가는 수험자의 성적 분포에는 신경을 쓰지 않고 수험자가 준거에 해당하는 또는 목표하는 능력을 갖추었는지 판단한다. 일상 생활에서 찾아볼 수 있는 가장 대표적인 준거 참조 평가는 운전 면허 시험일 것이다. 이러한 준거 참조 평가는 평가하고자 하는 능력을 가능한 한 직접 관측함으로써 이루어진다. 운전 면허 시험에서 수험자가 도로 주행시험에 통과하여야만 하듯이 영어 시험에서 수험자는 자신들의 영어능력을 보여줄 수 있는 과제를 수행하게 된다. 예를 들어 교사는 학생들에게 어떤 글을 읽고 소감을 발표를 시킨다거나 쓰도록 하고, 그들의 수행을 관찰하여 그들이 목표하는 언어 능력을 갖추었는지를 판단한다. [그림 3]은 이러한 수행 평가의 한 예이다. 학생들은 9명씩 한 조를 이루도록 나누어지며 한 학생에게 하나의 그림을 나누어준다. 각 학생들은 자신이 가진 그림을 설명하고 다른 사람의 그림에 대한 설명을 들음으로써 이야기를 순서에 맞게 구성하여 발표를 하거나 쓰도록 한다. 이러한 수행평가의 결과는 [그림 3]의 예에서 보듯이, 점수가 아니라 등급과 각 등급에 해당하는 기술자로 주어진다. 아래의 예에서 등급 1은 최상등급의 수행을 기술하며 등급 5는 최하의 등급을 기술한다. 각 등급의 기술자는 학습자들이 어떠한 언어 수행능력을 가지고 있다고 기술한다. 이와 같이 평가 결과가 학

생들의 능력을 직접 지시하면 몇 퍼센트의 학생들이 목표하는 언어 능력을 갖추었는지 판단을 할 수 있게 된다. 또 일정한 기준에 다다르지 못한 학생들에게 보충 학습을 제공할 수 있는 근거가 될 수 있다. 이러한 점은 7차 교육과정에서 도입된 단계 학습에 유용하게 적용될 수 있다.



(Gower and Walters, 1983에서 발췌)

과제 : 각 학생들은 그림을 한 장씩 가진다. 학생들은 자신의 그림을 다른 조원에게 설명하고 다른 사람들의 그림에 대한 설명을 들음으로써 이야기를 순서대로 구성한다.

등급

1. 자신의 그림을 쉽게 설명하고 다른 사람의 설명을 쉽게 이해한다.
2. 자신의 그림을 설명하는데 있어서 약간의 머뭇거림이나 가끔씩 문법적 오류를 보인다.
3. 자신의 그림을 설명하는데 자신감이 없어 보이며, 많은 문법적 오류를 보인다.
4. 자신의 그림을 설명하기 위해 단순히 단어를 나열하거나 몸짓에 심하게 의존한다.
5. 자신의 그림을 설명할 수 없다.

[그림 3] 수행 평가의 과제와 등급 기술자의 예

준거 참조 수행 평가의 또 다른 이점은 영어 의사소통 능력을 직접 측정할 수 있다는 것이다. 지금까지의 다지선다형 평가는 주로 읽기와 듣기 능력을 측정하는데 한정되어서 학습자의 말하기와 쓰기 등을 포함한 의사소통 능력을 측정하기에는 부적합하였다. 준거 참조 수행 평가의 도입은 또한 교실 수업에 많은 긍정적 파급 효과(washback effect)를 끼칠 수 있다. 학생들이 좋은 성적을 받는 것이 대학 입시에 절대적인 필수 조건으로 되어있는 우리의 교육상황에서 기존의 듣기와 읽기 기능만을 측정하던 다지선다형 시험에서 수행평가의 도입은 학생, 교사, 학부모의 관심을 언어의 모든 기능을 종합하는 영어 의사소통 능력에 보다 많은 관심을 가지게 되며, 교실 수업도 단순히 영어 문장을 읽고 번역하는 것이 아니라, 학생들의 의사소통 능력을 향상시킬 수 있는 활동으로 이루어지게 된다.

이렇게 수행 평가가 의사소통 능력을 향상시키고 측정하는데 유용할지라도 몇 가지 문제점을 가지고 있다. 가장 문제가 되는 것은 평가가 교사의 판단에 의존하기 때문에 공정성과 객관성이 결여될 우려가 있다. 이러한 문제를 해결하기 위해 영어 교사들의 재교육이 필요하다. 비록 교사들이 재교육을 받는다 하더라도, 그 기간이 충분치 못할 경우 오히려 교실 현장에서의 혼선만 초래할 우려가 있다. 이에 대비해 교사들이 직접 마련하기에는 부담이 되는 평가물(test material)과, 평가방법, 평가기준에 대한 구체적인 안내서를 교육부나 교사 위의 단계에서 준비를 해 주어야 한다. 또한 수행 평가가 제대로 이루어지기 위해서는 평가 시기의 변화가 필수적이다. 학교에서 행하여지던 지금까지의 평가를 보면 매학기의 성적이 중간시험과 기말시험의 결과의 합으로 산출되었다. 이때의 평가 내용은 한 학기동안 배운 내용의 대표적 표본으로 이루어진다. 하지만 이러한 고정된 평가시기를 수행평가에 적용하면 평가에서 측정할 대표적 표본 수행을 정하기도 어려울뿐더러 학생들의 수행이 단 한 번이나 두 번의 관찰에 의해 판단되기 때

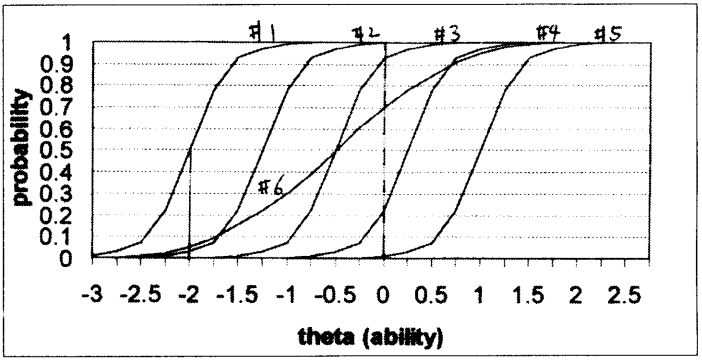
문에 정확하지 않을 수 있다. 학생들의 수행 능력을 보다 정확하게 관찰하기 위해서는 포트폴리오 방식의 평가 방법을 도입하여야 한다. 포트폴리오는 마치 개개 학생의 신상관리서와 같이 학생들 별로 수업 시간에 행하여진 영어 수행의 모든 자료를 수집하는 것이다. 이에는 학생들 스스로의 영어 능력에 대한 판단이나, 어떤 학생이 자신이 본 영화를 영어로 재미있게 이야기를 하였다던가 하는 일화나, 영어 경시대회에서 받은 점수 등 학생들의 수행에 대한 모든 정보들이 포함된다.

이렇게 수행 평가를 통해 학습자의 능력에 대해 직접적인 정보를 얻을 수 있음에도 불구하고, 수행평가를 교과 과정 정책에 대한 검증 목적으로 사용하기에는 한 가지 문제가 있다. 학습자의 수행을 직접 관찰하여야 하는 수행 평가는 대규모로 시행하기가 어렵다. 현재의 기술로는 대규모의 평가 시행시 직접 영어 수행을 관찰하여 측정해야 하는 말하기와 쓰기 능력은 측정하기 대단히 어렵기 때문에, 대규모의 영어 능력 시험은 듣기와 읽기 등의 수용적 언어 기능을 측정하는 다지선다형 평가로 제한될 수밖에 없다. 하지만 이러한 다지선다형 평가도 보다 유용하게 만들 수 있다. 그 한 방법으로 문항 반응 이론(Hambleton, Swaminathan, & Rogers, 1991)을 이용하여 평가를 구성하는 것이다. 문항 반응을 이용하면 고전 평가이론에서 평가 결과가 학습자간의 능력만 구분할 뿐 학습자가 어떠한 능력을 가지고 있는지 알려주지 못하는 문제점을 해결할 수 있다. 즉 어떤 점수가 대략 어느 정도의 능력과 연관이 있다고 가리킬 수 있다.

5.2. 문항 반응 이론

문항 반응 이론을 적용하여 만든 평가 문항은 고전 평가 이론에서의 문항과 근본적 차이를 보인다. 이전의 문항에서는 문항의 평이도와 변별도가 수험자 집단에 따라 상대적으로 결정되는 반면, 문항 반응

이론을 적용한 평가 문항은 문항의 평이도와 변별도가 평가가 실시되기 이전에 사전에 결정된다. [그림 4]를 예로 들어 살펴보자.



[그림 4] 문항 반응 이론의 문항 특성 곡선

[그림 4]는 평가의 문항 반응 이론에 기초하여 만들어진 평가 문항의 반응 곡선이다. 이는 문항 반응 이론을 이용한 평가가 특정한 수험자의 능력을 어떻게 일관된 점수로서 표시할 수 있는지 보여준다. X축은 수험자의 능력을 나타내며 일반적으로 -3에서 3의 범위로 표시하지만 이는 임의적이다. 6개의 곡선은 수험자들의 능력(X축)에 따라 각 문항에 대한 학생들의 반응을 나타낸다. Y축의 확률은 학생들의 능력에 따라 각 문항에 대해 올바른 대답을 할 확률이다. 이때 각 문항의 평이도는 해당 문항에 올바른 답을 할 확률이 50%인 학생들의 능력으로 표시한다. 예를 들어, 1번 문항에 대해 50%의 올바른 답을 할 확률을 가진 학생들이 능력은 -2이다. 즉 1번 문항의 난이도는 -2이다. 각 문항의 변별도는 문항 반응 곡선에서 50%에 해당하는 점의 접선의 기울기로 표시한다. 즉, 문항 반응 곡선의 기울기가 크다는 것은 그에 해당되는 능력을 지닌 학생들을 잘 변별한다는 것을 의미한다(문항 1의 경우는 -2의 능력에 해당되는 부분의 기울기가 가장 크

므로 이 문항은 -2의 능력을 가진 학습자들을 잘 변별한다고 할 수 있다. [그림 4]의 6개의 곡선 중 5 문항의 곡선은 난이도만 다를뿐 모두 같은 모양을 하고 있어 같은 변별도(접선들의 기울기가 같다)를 가지고 있다. 앞의 5 문항에 비해 문항 6의 곡선은 완만한 형태를 띄고 있다. 이 문항의 변별도는 다른 문항에 비해 낮지만 다른 문항보다 넓은 범위의 학생들을 구분한다.

이러한 문항 특성은 많은 문항 시험을 통해 평가가 시행되기 이전에 결정됨으로 문항의 특성이 시험을 치르는 집단의 특성에 영향을 받지 않는다. 시험에 대한 결과 점수는 학생들이 각 문항에 올바르게 답을 할 확률의 합과 같아진다. 예를 들어 0이라는 능력을 가진 학생의 점수는 1번과 2번 문항에서는 1의 확률을 가지며 3, 4, 5, 6에서는 각각 0.92, 0.2, 0, 0.7의 확률을 가진다. 이를 모두 합하면 3.82이 된다. 이와 같이 학생들의 능력에 대해 일정한 점수로서 표시함으로써, 점수로부터 학습자의 능력에 대한 직접적 정보를 얻을 수 있다.

고전 평가 이론에서 평가가 쉽다는 것은 많은 학생들이 많은 문항에 대해 올바른 답을 하여 높은 점수를 받은 것을 의미하지만 문항 반응 이론에서는 학생들이 평가가 쉽다고 느끼는 것은 낮은 능력의 학생들이 올바르게 답을 할 수 있는 문항들이 많이 출제되었다는 것을 의미한다. 그러나 학생들이 많은 문항에 답을 할 수 있다고 해서 학생들이 각 문항에 대해 올바르게 답할 확률이 변하는 것은 아니다. 문항 반응 이론은 이러한 확률을 점수로 표시하는 것이다.

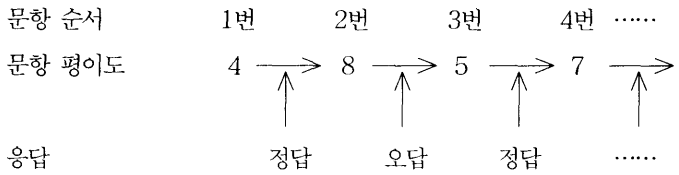
문항 반응 이론을 이용한 표준화 평가는 기존의 표준화 평가에 비해 많은 이점이 있다. 우선, 두 개 이상의 평가를 연결할 수 있다. 예를 들어, 학생들의 영어 능력의 향상 정도를 일정 기간 측정한다고 하자. 이러한 경우 학생들의 수행능력의 변화를 관찰하기 위해 실험 초기와 말기에 학생들의 언어 능력을 측정하고 때로는 중간단계에서도 학생들의 수행능력을 측정한다. 이와 같이 같은 집단의 학생들을 여러

번 측정할 때는 같은 시험을 여러 번 이용할 수 없다. 따라서, 서로 다른 몇 개의 시험을 이용하여 학생들의 언어 능력의 향상 정도를 측정하여야 한다. 이때 언어 능력의 향상 정도를 측정하기 위해 평가 결과를 전체 문항 중에서 몇 퍼센트나 맞추었는지를 비교하는 것은 의미가 없다. 나중에 본 시험이 난이도가 높으면 학생들의 실력이 향상되지 않고 오히려 저하된 것으로 나타날 수 있기 때문이다. 문항 반응 이론을 이용하게 되면 평가들간의 난이도를 고려하여 시험 결과들을 하나의 척도에 놓고 비교할 수 있다.

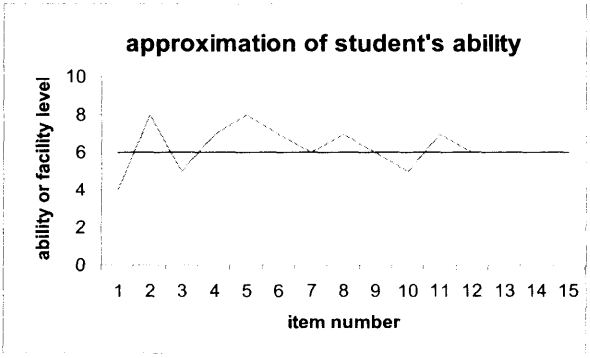
문항 반응 이론을 이용하면 평가의 난이도를 조정하는 것이 가능하기 때문에 서로 다른 시기에 시행되는 평가들의 난이도를 일정하게 조절할 수 있다. 즉, 학생들이 시험을 3월에 치르건 8월에 치르건 학생들에게 동등한 난이도 수준의 평가를 제공할 수 있다. 고전 평가 이론에 기초한 표준화 평가에서는 평가의 난이도가 평가가 시행된 후에 평가를 치른 수험자가 누구냐에 따라 상대적으로 결정되기 때문에 평가들 간의 난이도를 비교하는 것이 대단히 어려웠을 뿐만 아니라 서로 다른 평가들로부터의 결과를 비교하기가 대단히 어려웠다.

문항 반응 이론의 가장 큰 장점은 컴퓨터 적응 평가(computer adaptive test)를 가능케 하는 것이다. 컴퓨터 적응 평가는 문제 은행으로부터 학습자의 수준에 적합한 문항들을 선택함으로써 학습자의 수준을 찾아준다. 이는 대규모의 문제 은행의 구성이 필수적이며, 각 문항들의 학습자의 수준에 따른 반응 특성이 평가 시행 이전에 결정된다. 각 문항에 대한 학습자의 반응에 따라 학습자에게 제시될 다음 문항이 결정된다. 예를 들어, [그림 5]에서처럼, 문제 은행의 문항들이 0에서 10이라는 능력을 측정한다고 가정하고, 컴퓨터가 학습자에게 4라는 평이도(4의 능력을 갖춘 학습자가 이 문항에 올바른 답을 할 확률이 50%이다)를 나타내는 문항을 제시하였다고 하자. 학습자가 이 문항에 대한 올바른 답을 제시하면, 다음 문항으로서 컴퓨터는 4보다

높은 수준의 문항을 선택하여 제시한다. 만약 학습자가 올바른 답을 하지 못하면 컴퓨터는 5보다 낮은 수준의 문항을 선택한다. 세 번째 문항에 대한 선택도 이와 같은 방식으로 학습자가 올바르게 답을 하는지의 여부에 따라 두 번째 문항보다 높은 수준(보다 어려운)의 문항이 주어지거나 보다 낮은 수준(보다 쉬운)의 문항이 주어지게 된다. [그림 6]은 컴퓨터 적응 검사가 학습자의 반응에 따라 학습자의 수준에 문항의 선택을 접근시켜 나아가는 과정의 예이다. 이 예에서의 학습자는 6의 능력을 가지고 있으며 컴퓨터는 4, 8, 5, 7, 8, 6, 7, 6, 5, 7, 6, 6 수준의 문제들을 순서대로 제시하고 있다. 평가가 진행되어가면



[그림 5] 수험자의 반응에 따른 컴퓨터 적응 평가의 문제 제시과정의 예



[그림 6] 컴퓨터 적응 평가가 문항의 난이도를 수험자의 능력에 맞추어 나가는 과정의 예

서 문항 수준이 학습자의 수준에 보다 근접하여 간다. 이러한 컴퓨터 적응 평가를 이용하면 수험자가 원하는 시기에 아주 적은 수의 문항으로 빠르게 수험자의 능력을 정확하게 측정할 수 있다.

문항 반응 이론을 이용한 표준화 평가는 고전 평가 이론에 기초한 표준화 평가가 수험자 집단의 특성에 따라 결과 해석이 달라지고, 같은 목적의 평가들 간의 결과를 비교할 수 없었던 단점을 극복할 수 있다. 특히 학생들의 능력을 객관적이고 직접적으로 해석할 수 있는 점은 교과과정 정책의 시행과 검증에 대단히 유용하다고 할 수 있다. 하지만 이의 시행을 위해서는 많은 문제점들이 해결되어야 한다. 문항 반응 이론에 기초한 평가를 만들기 위해서는 문항들의 특성과 평가의 신뢰도 타당도 등을 유지하기 위한 많은 시간과 인력이 필요하므로, 평가를 담당할 전문 연구기관의 결성이 필요하다. 이러한 연구기관에 의해 평가가 시행되기에 앞서 많은 검증과정을 거쳐야만, 누구나 결과를 신뢰할 수 있는 평가를 만들 수 있다.

이러한 표준화 평가는 대개 광범위한 언어 능력을 측정하지만, 중학생과 고등학생에게 요구되는 영어 능력 수준을 감안할 때, 단일한 평가로서 모든 학생들을 평가하는 것은 바람직하지 않을 것이다. 문항 반응 이론을 이용한 표준화 평가는 일반적으로 가장 효과적으로 학생들의 능력을 구분할 수 있는(학생들의 능력에 대해 가장 정확한 정보를 얻을 수 있는) 영역을 목표에 두고 만들어진다. 이러한 영역은 평가의 목적에 따라 다르게 조정된다. 예를 들어, 장학생을 선발할 목적이면 평가에 높은 수준의 문항을 많이 포함시켜 높은 수준의 영역에서 보다 정확한 정보를 얻도록 만들어진다. 중학생과 고등학생을 단일한 평가로 측정하게 되면 평가가 측정하고자 하는 목표 영역이 지나치게 넓어지게 되어 평가구성과 결과해석의 효율성을 떨어뜨리게 된다. 따라서 두 집단의 학생들에게 맞추어진 평가를 따로 운영하는 것

이 바람직할 것이다.

지금까지 설명한 표준화 평가는 대규모로 시행할 수 있으며, 평가 결과가 학습자의 능력을 직접적으로 가리킬 수 있기 때문에 교육과정 목표의 달성여부를 판단하는데 쉽게 이용될 수 있다. 하지만 문항 반응 이론에 기초한 평가라 할지라도 이전의 다지선다형 평가처럼 읽기와 듣기 등의 수용적 언어 능력만을 측정하는 한계를 가진다. 따라서, 일선 교실에서는 학생들의 영어 능력을 정확하게 측정하기 위해서는 표준화 평가의 결과뿐만 아니라 말하기와 쓰기 능력을 측정하는 수행평가를 같이 이용하여야 한다.

6. 결론

지금까지 이 연구는 평가의 결과가 교과과정 목표의 달성여부를 결정할 근거로 이용될 수 있는가의 관점에서, 아주 오랜 동안 우리들의 머리 속에서 시험이라고 하면 대표적으로 떠오르는 형태인 다지선다형 평가의 문제점들과 그 개선 방향을 찾아보았다. 평가를 교과과정 정책의 목표수립과 시행에 이용하기 위해서는 평가의 결과가 학습자들의 능력을 상대적이 아닌 절대적인 척도 상에서 나타내야 한다. 지금까지 교사와 평가기관에서 만들어져온 다지선다형 평가로부터는 학습자의 능력에 대한 직접적인 정보를 얻을 수 없었다. 이에 대한 대안으로 일반 교실에서는 수행평가를 이용하여 학생들의 영어 수행능력을 직접 관찰하여 측정할 것을 제안하였다. 그러나 이러한 수행평가는 시행상의 제약으로 인해 대규모의 시행이 어려우므로, 교실을 벗어나 전체학생을 평가할 시는, 문항 반응 이론에 기초하여 표준화 평가를 구성할 것을 제안하였다.

이러한 평가 개혁을 유도하기 위해서는, 교육부는 일선 교사들을 재

교육하여야 하며, 교사들이 이용할 수 있는 구체적인 평가 도구(평가 방식과 내용)를 마련하여야 한다. 또한 문항 반응 이론에 기초한 표준화 평가를 누구나 신뢰하고 이용할 수 있도록 하기 위해서는 평가 연구 전문기관을 설립하여 이 기관으로 하여금 표준화 평가를 담당하도록 하는 것이 필요할 것이다.

이러한 평가의 변혁은, 특히 교실에서의 수행평가의 도입은, 수업내용과 방식에 있어서 많은 변화를 가져올 것이다. 지금까지처럼 다지선다형 평가에 의존하였을 때에는 학생들의 주된 학습 목표가 읽기평가나 듣기평가에서 높은 점수를 받는 것이었다. 따라서 언어 산출기능인 말하기나 쓰기 기능은 상대적으로 소홀히 되어왔다. 하지만 네 가지 언어 기능을 고르게 측정하는 수행 평가의 도입은 교사와 학생들의 학습 목표를 긍정적인 방향으로 바꾸어 놓을 수 있다.

참고문헌

교육부, 『외국어과 교육 과정(I)』, 대한교과서주식회사, 1998.

한국교육과정평가원, 『중·고등학교 영어평가 도구·개발 지원방안 연구』, 출판예정

Davidson, F., *EIL 360 lecture note*, Unpublished manuscript, University of Illinois at Urbana-Champaign, 1993.

Hambleton, R., Swaminathan, H., & Rogers, H., *Fundamentals of item response theory*, Newbury Park, CA: SAGE Publications, 1991.

McNamara, T., *Language testing*, Oxford, UK: Oxford University Press, 2000.

Peirce, B. N., Demystifying the TOEFL reading test. *TESOL Quarterly*, 26, 665-689, 1992.

Gower, R & S. Walters., *Teaching Practice Handbook: A reference book for EFL teachers in training*, Oxford: Heinemann, 1983.

강성우(Sung-Woo Kang)
청주교육대학교 영어교육과
주소 : (361-712) 충북 청주시 수곡동 135
청주대학교 영어교육과
Tel : 043-299-0821
E-mail : skang4@chollian.net