교육 평가의 혁신을 위한 테크놀러지의 활용

박 주 용 (세종대 교육학과)
jpark@sejong.ac.kr

Applying Information and Communication Technology for Advancing Educational Assessment

Jooyong Park (Dept. of Education, Sejong University)

요 약

평가는 교수와 함께 교육 목표를 이루는데 있어 핵심적인 역할을 한다. 그렇지만 평가는 교수법의 발전 속도에 비교하면 상당히 느릴 뿐만 아니라, 교사는 물론 학생들에게 부정적으로 비추어지고 있다. 본 논문에서는 시험이 학생들을 동기화시킬 뿐만 아니라 그 자체로 중요한 학습 경험임을 강조하면서, 학습을 위한 평가를 활성화시키기 위해 테크놀러지를 이용한 평가 방식들을 소개하였다. 개발은 물론 어느 정도 연구가 이루어진 테크놀러지를 활용한 평가 기법을, 수행평가와 선다형의 개선안으로 대별하였다. 수행평가를 위한 기법에서는 논술채점, 지식 지도 제작법, 그리고 학습자 모형을 이용한 평가기법이, 선다형의 개선안으로 적응형 컴퓨터화 검사, 다중평가 및 변형선다형 방식이 각각 상술되었다. 결론에서는 평가를 통한 교육 효과를 극대화하기 위해서는, 교사는 물론 학생들의 평가에 대한 태도 변화 교육과 실제 교육 현장에서 쉽게 활용될 수 있는 컴퓨터를 이용한 평가 기법의 개발과 보급의 중요성이 강조되었다.

주제어: 교육 평가, 컴퓨터화 검사

I. 서론

시험이 주는 스트레스를 생각하면, 시험이 없는 학교를 꿈꾸는 학생들이 있다는 것이 놀라운 일이 아니다. 시험 상황이 야기하는 불안, 경쟁심, 게다가 결과가 비교될 때 느끼는 열등감과 패배의식 등은 대부분의 학생들에게 시험을 피하고 싶은 마음이 들게 하기에 충분하다. 시험이 달갑지 않은 것은 교사들에게도 마찬가지이다. 성적을 매겨야하고 또 열심히 공부하게 하려고 시험을 시행하지만, 문제를 만들어야 하고 채점을 하는 것은 부담스럽기 때문이다.

그렇지만 시험이 없는 교육과 훈련을 상상하기 어렵다. 시간이나 그 밖의 자원이 제한된 상황에서 일부만을 선발해야 하고 또 교육과 훈련 효과를 높이기 위해서는 시험과 같은 평가 절차가 필요하다. 학교를 졸업하고 나서도 입사승진 시험으로 이어지면서 '평생 수험생' 으로 사는 대부분의 한국인들에게 있어, 시험은 학습을 위한 도움이로서보다는 통과해야 할 시련으로 우리의 삶 속에 깊숙하게 자리 잡고 있다. 특히 대부분의 학생들은 삶을 위한 시험으로서가 아니라 시험을 위한 삶으로 학창시절을 보낸다. 그런데 정말 시험이나 평가는 필요악일까? 교육에서 평가가 차지하는 역할은 당락을 결정하거나 서열을 정하는 일뿐일까? 물론 아니다. 평가에 대한 부정적인 이미지는 많은 경우 평가가 교육을 위해 사용되기보다는 행정적 용도로 사용되어졌기 때문이다. 평가는 주로 선발이나 학교간 비교를 위해 활용되어져 왔다. 우리나라의 경우 내신이 중요하기는 하지만 좋은 대학에 들어가려면 대입 수능시험에서 높은 점수를 받아야 한다. 미국의 경우 주 단위의 표준화된 대규모 시험 결과에 따라 잘하면 지원금을 지급하지만 못했을 경우 교사나 교육 관료가 해임되거나 학교의 존립에까지 영향을 준다. 이 때문에 점수를 높이는 것이 중요하였다. 문제는 시험 점수를 높이기 위한 교육이 결과적으로 교육의 질을 오히려 떨어뜨린다는 점이다 (예, Amrein & Berlinger, 2002; Linn, 2000). 이 연구자들은 이런 중요한 대규모 시험의 비중을 낮추고 교실 장면에서의 형성 평가를 활성화할 것을 주장하고 있다 (예, Black & Wiliam, 1998).

형성 평가는 평가의 학습적 측면을 강조한다. 평가가 학습을 촉진하는 대표적인 예는 시험 효과에서 볼 수 있다 (Foos & Fisher, 1988; Glover, 1989; Park, in press). 시험 효과란 학습한 내용에 대해 중간에 시험을 본 집단이 같은 시간동안 복습을 한 집단보다 나중의 시험에서 더 높은 점수를 얻는 현상을 가리킨다. Glover는 또한 중간에 보는 시험을 선다형으로 볼 때와 단답식으로 볼 때의 차이를 비교하였는데 단답식으로 볼 경우에 더 큰 시험 효과를 관찰하였다. 요컨대 인출 부담이 클수록 나중에 그 내용을 더 잘 기억한다는 것이다

형성 평가의 중요성에 대한 이같은 인식에도 불구하고 실제 교실 장면에서의 학습을 위한 평가는 여전히 활성화되고 있지 않

고 있다 (예,Earl, 2003). 그 한 이유는 평가에 따른 교사의 노력과 시간 때문이다. 최근 이런 문제를 해결하면서 평가의 효율성을 높이기 위한 새로운 기법들이 개발되고 있다. 본 논문에서는 이들을 테크놀러지를 활용한 구성형 검사 기법들과 선다형의 개선 기법들로 나누어 소개하고자 한다.

II. 테크놀러지를 활용한 구성형 평가 기법들

1. 자동화된 논술 채점 시스템

논술은 가장 오래된 평가 방법이라 불릴 수 있다. 중국의 과거 시험이 전형적인 예로 대개 하나의 주제에 대해 기본적인 사실을 바탕으로 어떤 논리 정연한 주장이 펼쳐진다. 추론을 배제할 수 있고 단편적인 지식의 나열이 아닌, 관련된 지식이 어느 정도로 체제화되어 있는지는 물론, 고차적인 추론 능력을 평가하는 최상의 도구라 할 수 있다.

이런 장점에도 불구하고 논술은, 학급의 규모가 커짐에 따라 회피되어 왔다. 그 이유는 평가에 드는 시간과 노력이 엄청나기 때문이다. 여기에다 만일 여러 사람이 나누어 채점할 경우 채점자의 주관이 개입되므로 결과의 공정성도 문제가 될 수 있다. 이런 문제를 해결할 수 있는 자동화된 논술 채점 시스템에 대한 연구가 꾸준히 이루어져 최근 실제 학교장면에서 사용될 수 있을 정도의 성능을 갖춘 상용 프로그램들이 등장하였다.

이들 프로그램은 실제로 사람들에게 논술문을 채점하는 방법을 가르치듯 컴퓨터에게 특정한 기법을 가르친 다음 이것을 활용하여 채점하도록 하는 것이다. 보다 구체적으로 보면, 먼저 전문가들이 특정한 형식에 따라 채점한 내용을 컴퓨터에 입력시킨다. 그 다음 방대한 양의 채점된 내용을 통계적 기법을 사용하여 단순화하고, 이를 활용하여 새로운 논술문을 채점한다.

가장 오래된 논술채점 프로그램 중의 하나는 Page에 의해 제안된 PEG(project essay grading)이다 (Page, 1966; Page & Petersen, 1995). Page는 사람들이 논술문을 평가할 때 유창성, 복잡성, 어법 등과 같은 주요 내재적 변인들인 trin[1]에 근거한다고 본다. 컴퓨터가 trin을 활용할 수 없음은 자명하다. 그렇지만 trin에 준하는 몇 몇 형식적 특성들인 proxes를 사용할 수 있다. 유창성과 상관이 있는 논술문의 길이, 복잡성과 관련된 전치사의 수, 관계 대명사의 사용정도, 그리고 어법에 대응되는 단어 길이의 변화 등이 그 예이다. Page는 한 주제에 대해 잘 쓰여진 여러 글에서 나타나는 proxes들을 찾아낸 다음, 이들을 독립변인으로 하고 평가를 종속 변인으로 하는 중다 회귀모형을 설정하였다. 실제 평가는 이 모형에 구해진 계수들에 채점될 개별 논문의 proxes 값들을 대입하여 얻은 추정치를 구하는 것이다. 단어나 내용에 대한 고려가 전혀 없이 논술문의 표면적 특징만을 가지고도, 두 명으로 이루어진 채점자 집단간 상관보다 채점자집단과 PEG간의 상관이 더 높다는 결과를 얻었다 (Page & Petersen, 1995).

전술한 것처럼 PEG의 경우 단어나 내용에 대한 고려가 전혀 없다. 그렇지만 잠재적 의미분석 (latent semantic analysis: LSA)에서는 벡터-공간을 이용하여 주요 문건과 그 문건에 포함되어 있는 단어(혹은 개념)을 표상한다 (Landauer, Foltz & Laham, 1998; Thompson, 1998). 어떤 한 주제와 관련된 문건의 수는 물론 그 문건에 포함된 단어의 수는 엄청나게 많다. 예를 들면 "사학비리" 와 관련된 문건으로는 신문, 교육인적자원부의 조사 보고서, 그리고 이와 관련된 논문 등이 있을 수 있다. 이제 이들 가운데 500개를 선정한 다음, 그 문건들에서 사용된 내용 단어를 헤아려 보니 10000개였다고 가정해보자. 각 단어는 10000개의 행에 각 문건은 500개의 열에 차례로 입력한 다음 각 문건에서 각 단어의 유무를 1과 0을 이용하여 나타내면 10000*500의 행렬이 만들어진다. LSA는 이 방대한 행렬을 SVD(singluar value decomposition)[2]이라는 기법을 통해 분해한 다음, 이 중, 예를 들면 100개의 차원만을 사용하는 축약된 새로운 행렬을 만들어 낸다. 흥미로운 점은 이 행렬은 자주 나오지 않는 단어들을 일종의 "오염 요인" 으로 간주하고 이를 제거하면서 관련이 있는 단어들끼리의 상관을 높여준다는 것이다. 그 결과 동시에 같이 나타난 적이 없는 단어들도 다른 단어와의 높은 발생빈도로 인해 높은 상관이, 따라서 의미적 연관성이 있음이 발견되기도 한다.

이제 LSA를 논술의 채점으로 확장하는 것은 간단해진다. 문건을 미리 채점된 다른 논술로 대체한 다음 채점될 논술문에서 사용된 단어 행에는 1 사용되지 않은 단어 행에는 0을 입력하여 벡터를 만든다. 이 벡터와 SVD를 이용해 만들어진 새로운 행렬의 각 열 벡터와의 유사성은 코사인 계수[3]를 통해 간단히 계산되는데, 이 계수에 가장 가까운 문건의 점수를 할당하면 된다.

ETS에서 개발 활용되고 있는 e-rater는 PEG에서처럼 선형 회귀분석을 사용한다 (Burstein, Chodorow, & Leacock, 2003). 하지만 LSA에서와 같은 의미적인 특성도 측정된다. 이 특징들은 통사적 다양성, 논술 (discourse) 분석, 주제 내용, 그리고 어휘 복잡성 등으로 대별된다. 이들은 다시 50여개의 특성들로 세분되지만, 이 가운데 20개 정도면 회귀모형을 찾아낼 수 있다고 한다. 나머지 과정은 PEG에서와 동일하다.

한편 베이즈 접근법(Bayesian approach)을 이용한 채점기법도 개발되고 있다 (Rudner, 2002; Rudner & Liang, 2002). 이 기법은 애당초 자료를 성공/실패 혹은 초급/중급/고급의 구분에서처럼 유한한 수의 범주로 구분하는 결정이론을 논술 평가에 적용한 것이다. 어떤 학생의 논술에서 나타난 특성의 벡터를 **z**라 할 때 이 학생이 예를 들어 고급 수준에 있을 확률은 베이스정리를 이용해 다음과 같이 기술될 수 있다:

$$P(\text{고급}|\mathbf{z}) = \frac{P(\mathbf{z}|\text{고급})*P(\text{고급})}{P(\mathbf{z}|\text{초급})*P(\text{초급})+P(\mathbf{z}|\text{중급})*P(\text{중급})+P(\mathbf{z}|\text{고급})*P(\text{고급})}$$

다른 수준 즉 중급 및 초급 수준에 대한 조건부 확률도 분자에 있는 조건부확률과 사후확률을 바꾸면 계산할 수 있다. 이제 목

1) trin과 다음에 나오는 prox는 각각 영어의 intrinsic 과 approximation으로부터 만들어진 용어이다.

2) m*n으로 이루어진 행렬 A는 3개의 다른 행렬인 U ΛV'의 곱으로 분해될 수 있는데 U는 행 벡터이고 Λ는 K개의 대각행렬이며 V'는 전위된 열 벡터이다.

3) $\cos\Theta = \frac{\mathbf{v}\cdot\mathbf{w}}{||\mathbf{v}||\ ||\mathbf{w}||}$

표는 이런 논술에서 나타난 세부특징 벡터를 이용하여 어떤 수준에 속할지를 판단하면 된다. 그런데 사전확률을 계산하려면 반응의 분포에 대한 가정이 필요하다. 문서 분류에서는 흔히 다변인 베르누이 (multivariate Bernoulli) 모형과 다항 (multinomial) 분포 모형이 사용된다. 전자는 각 논술문을 계산된 모든 세부특징의 특별한 사례로 간주하고, 세부특징의 존재 유무를 중심으로 조건부확률을 계산한다. 후자에서는 각 논술문을 계산된 세부특징의 한 샘플로 간주하여 세부특징이 여러 번 사용되는 것을 고려할 수 있다. 이들 모델과는 독립적으로 조건부확률간의 우도에 대해 결정을 내리는 방법 역시 최대우도법, 최소 확률오차법, 최대사후확률법, 및 베이즈 위험감수법 등으로 다양하다 (Rudner, 2002). 더욱이 조건부 확률 계산에서 구체적으로 사용되는 세부특징은 단어나 구절 혹은 특정 단어가 다른 특정한 단어보다 먼저 나옴을 나타내주는 논항이 될 수도 있고, 단어들도 어근만을 사용할지 아니면 모든 단어를 사용할지 등의 차이가 있을 수 있다. Rudner와 Liang (2002)은 이들 변인에 따른 차이를 알아보기 위해 고등학교 학생들이 쓴 75단어 정도 되는 생물학관련 논술문을 분석하였다. 그 결과 베르누이 모형이 다항 모형보다 정확하며, 단어나 구보다 논항이, 그리고 모든 단어를 고려한 경우가 어근만 사용한 경우보다 더 정확하다는 것을 발견하였다. Rudner와 Liang은 위에서 언급된 회귀모형이나 잠재적 의미 분석에 비해, 베이즈접근에서는 훨씬 더 소수의 자료를 가지고도 정확한 결정을 내릴 수 있다는 점을 강조한다. 더구나 특정 단어나 구의 유무만을 가지고 판단하는 다른 모형과는 달리 논항 즉 단어나 구가 논술문에서 순서까지도 고려할 때 채점이 정확하다는 점을 밝힌 점도 중요한 발전이라 할 수 있다. 앞으로의 연구는 범주의 수가 증가함에 따라 그 정확도가 여전히 유지될 수 있는지는 밝히는 일이겠다.

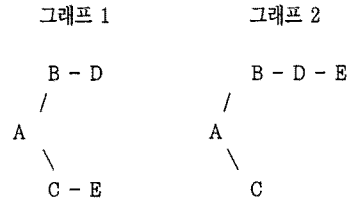
2. 지식 지도제작법(knowledge mapper)을 통한 지식의 구조화

기억 연구자들은 개념적 지식을 표상하기 위해 오래 전부터 개념도(concept map)를 사용하여 왔다 (예, Anderson, 1983). 개념도란, 신경망처럼, 개념을 표상하는 마디(node)와 개념들을 연결해주는 연결선(link)으로 구성되어 있다. 지식 지도제작법은 이 개념도를 교수는 물론 평가도구로 활용한 기법이다. 즉, 교사들은 특정 주제와 관련된 주요 개념과 그들의 관계를 개념도를 통해 요약 정리해 줄 수 있을 뿐만 아니라 학생들에게 이런 개념도를 그리게 한 다음 이를 평가할 수도 있다.

지식 연결도식은, 표준화된 선다형 검사가 포착하기 어려운 총체적 지식 구조를 표상할 수 있을 뿐만 아니라 주요 개념을 도해하기 때문에 언어 구사 능력에 구애되지 않으면서 이해 정도를 표현할 수 있다는 장점이 있다. 이 방식을 평가 장면에 직접 활용할 수 있도록 하는 컴퓨터 프로그램을 구현하는 동시에 그 타당성을 다각적으로 검토하는 연구는 CRESST에서 이루어지고 있다 (예, Chung, Baker, & Cheak, 2002; Herl, Baker, & Niemi, 1996; Herl, O'Neil, Chung, & Schacter, 1999; Ruiz-Primo, Schultz, Li, & Shavelson, 1999; Ruiz-Primo, Shavelson, & Schultz, 1997; Schacter, Herl, Chung, Dennis, & O'Neil, 1999).

일반적인 평가절차는 먼저 주제 영역을 정한 다음, 전문가에게 그 주제와 관련된 중요한 개념과 그들 간의 관계를 도해하게 한 다음, 이를 준거로 실제 학생들의 도해를 채점하도록 한다. 하지만 이 일반적 절차 내에서도 다양한 변형이 가능하다. Ruiz-Primo 등(1999)에 따르면, 개념 마디나 연결 관계 등을 평가자가 제공하고 학생들이 그것을 채우게 할 수도 있고 아니면 학생이 모두 구성하게 할 수도 있다. 학생들이 도해한 것을 평가하는 방법도 여러 가지 이다. 예를 들면, 전문가가 도해한 것과 정확히 일치하는 것만 점수를 줄 수도 있고, 전체 반응 중 전문가가 도해한 것과의 일치 비율을 사용할 수도 있다 (Ruiz-Primo 등, 1997). 그렇지만 일반적으로 채점에서 고려되는 변인은 의미적 내용, 구성 (organized structure), 사용된 용어수, 연결마디의 수, 그리고 용어에 대한 평균 연결마디의 수 등이다. 의미적 내용은 의미적 연결의 일치도를 측정하고 구성은 연결망 내에서 인접 용어들간의 유사성의 정도를 측정한다. 이를 위해 주로 사용되는 척도는 Goldsmith, Johnson 그리고 Acton(1991)이 제안한 C가 사용된다. C척도는 각 마디에 대해 두 도해간에 교집합과 합집합을 구한 다음 전자를 후자로 나누어 계수를 구한다. 이렇게 구해진 계수의 총합을 전체마디수로 나눈 값이 C인데 전혀 상관이 없는 0에서 완전히 똑같은 1까지의 범위를 갖는다. Goldsmith 등이 제시한 예를 더 단순화시킨 예는 그림 1에 제시되어 있고 C의 계산 과정은 그림 2에 제시되어 있다.

그림 2의 마지막 열에 개별 마디로부터의 계수가 정리되어 있는데, 이들 계수의 합은 3이고 마디의 수는 5개이므로 C=3/5=.6이 된다. 참고로 이 두 그래프간의 상관을 구하기 위해, 각 마디쌍으로 이루어진 거리를 나타내면 그림 3과 같은데 그 값은 0이다.



[그림 1] 5 마디로 이루어진 두 개념도

	이웃 마디		교집합		합집합		
마디	그래프1	그래프2	집합	크기	집합	크기	계수
A	{B,C}	{B,C}	{B,C}	2	{B,C}	2	2/2
B	{A,D}	{A,D}	{A,D}	2	{A,D}	2	2/2
C	{A,E}	{A}	{A}	1	{A,E}	2	1/2
D	{B}	{B,E}	{B}	1	{B,E}	2	1/2
E	{C}	{D}	{}	0	{C,D}	2	0/2

[그림 2] 그림 1의 개념도로부터의 C 도출 방법

	A	B	C	D	E
A	-	1	1	2	2
B		-	2	1	3
C			-	3	1
D				-	4

	A	B	C	D	E
A	-	1	1	2	3
B		-	2	1	2
C			-	3	4
D				-	1

[그림 3] 그림 1의 개념도에서 각 마디쌍 간의 거리를 나타낸 매트릭스

개념도에 대한 검사로서의 신뢰도와 타당도에 대한 다양한 연구가 진행되고 있는데, 다른 검사와의 상관은 중간 정도이나 사용 가능성에 대한 교사들의 평가는 긍정적이다 (예, Chung 등, 2002).

3. 문제 해결 과정에 대한 학습자 모형을 활용한 평가

지금까지 살펴본 논술이나 지식 개념도는, 지필식에서처럼 최종 산물의 평가에 그 초점이 있다. 이 절에서 살펴볼 복잡한 문제해결은 최종 산물뿐만 아니라 과정에 대한 정보도 얻을 수 있고 따라서 교수과정과도 밀접한 관련이 있다. 교수자(tutor)로서의 컴퓨터는, 일련의 순서로 진행되는 교수, 평가와 피드백이 반복되는 행동주의적 교수 공학 기법을 쉽게 구현할 수 있었다 (Burton, Moore, & Magliaro, 1996). 여기에 인공지능의 원리와 기법이 더해진 지능형 교수 시스템은, 학생들의 수행 수준에 따른 더 구체적이고 적절한 피드백을 제공할 수 있게 되었다. 그렇지만 교수와 평가가 순환적으로 하나의 고리를 형성하고 있어, 지능형 교수 시스템을 평가 도구로 전환하는 것은 강조하는 점만을 달리하는 것 일 뿐이다 (예, Ager, 1990; Frederiksen & White, 1989).

지능형 교수 시스템은 여러 하위 체계로 구성되지만 가장 중요한 것은 학습자 모형이다 (Greer & McCalla, 1994). 학습자 모형은 학습자의 수행을 모니터링하고 이를 바탕으로 현재의 지식 수준을 추리하며 또한 어떤 문제에 대한 수행 수준을 예측하는 역할을 담당한다. 문제는 학습자 모형을 구축하는 일이 쉽지 않다는 점이다. 학습자 모형은 주로 인지적 과제 분석(cognitive task analysis)을 통해 이루어지지만, 최근 경험적 자료를 자기조직화하는 (self-organizing) 방법도 활발해지고 있다.

인지적 과제 분석 (cognitive task analysis)의 주요 목표는 한 영역의 전문가로부터 그 영역에 관한 해박한 지식을 알아내 이를 모형으로 만들어 내는 것이다 (이에 대한 주요 기법을 보려면 Jonassen, Tessmer, & Hannum, 1999를 참조하시오). 이렇게 만들어진 전문가 모형과 학습자의 현재 모형과의 차이를 줄이기 위한 일련의 절차는 교수과정이고, 평가는 학습자의 현재 모형을 찾는 과정이라 할 수 있다. 인지적 과제 분석 기법의 하나인 PARI (Precursor, Action, Result, & Interpretation의 머리글자)는 전문가가 팀을 이루어 실제적인 상황에서 서로에게 문제를 내고 이를 풀게 한 다음 그 과정에 대해 다시 토의를 거쳐 최종적인 전문가의 해법을 만들어낸다. 그 다음 초보자나 중간 수준의 문제 해결과정을 찾아내고 이들을 교수나 평가를 위해 활용한다. 실제로 이 방법은 여러 장면에서 활용되고 있다. F-15 전투기의 격납고에 있는 수압조절장치의 고장 수리 기술을 비디오 이미지를 통해 교육하는 시스템인 HYDRIVE가 그 하나고 (Mislevy & Gitomer, 1996), 치위생사를 위한 자격시험 및 계속 교육용 컴퓨터화 평가 모듈의 개발이 또 다른 예이다 (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999).

인지적 과제 분석이 지능형 교수 시스템에서 중요한 역할을 할 것이지만, 더 광범위하게 사용되기에는 어려운 문제점이 있다. 하나는 전문가들이 참여가 꼭 필요하다는 것과 분석 자체가 시간이 오래 걸린다는 점이다. 이 때문에 어떤 경우에는 연구자가 자신의 직관에 근거하여 임의적으로 설정하기도 한다. 이보다 더 큰 문제는 전문가의 문제해결과정과 초보자의 문제 해결과정에 질적인 차이가 있어, 전문가 모형을 바탕으로 초보자의 문제 해결 능력을 측정하기 어렵다는 점이다(예, Lesh & Kelly, 1996). 이상의 문제를 해결하는 한 방법은 자료에 근거하여 경험적으로 학습자 모형을 찾는 방법이다 (Harp, Samad, & Villano, 1995; Hurst, Casillas, & Stevens, 1997; Stevens, keda, Casillas, Palacio-Cayetano, & Clyman, 1999; Vendlinski & Stevens, 2000). IMMEX project는 이 방법을 통한 평가의 특성을 다각적으로 연구하고 있다. IMMEX는 상호작용하는 멀티미디어를 통한 훈련(Interactive Multi-Media Exercises)의 머릿글자이다. 이 시스템은 문제 해결과 관련된 지식이나 검사 정보를 메뉴에서 찾아볼 수 있는 환경에서, 문제의 제시에서 시작하여 결론에 도달하는 동안 학생이 어떤 메뉴를 어떤 순서로 찾아보았는지를 기록한다. 예를 들어 어떤 미지의 물질의 정체를 확인하는 문제에서 적색 리트머스 검사, 황산에 대한 반응 등이 메뉴판에 있어 학생들은 어떤 검사가 필요한지를 스스로 판단하여 이들 검사를 원하는 만큼 실시해볼 수 있다. 각 학생이 이들 (예를 들어 10개의) 메뉴 중 어느 것을 어떤 순서로 사용하였는지에 대한 정보는 컴퓨터에 기록된다. 이 정보는 검색 경로로 도해 (search path mapping) 될 수 있는데, 화살표를 이용하여 그림으로 나타낼 수도 있고, 그림 4에서와 같이 숫자를 이용하여 나타낼 수도 있다:

여기서 0은 참조되지 않은 메뉴임을 알려주며, 메뉴 11은 답을 제시하는 단계를 나타낸다. 따라서 학생 001은 10번에서 시작하여 3, 4, 5, 6번 메뉴를 참조한 다음 답을 제시하였고, 학생 002는 9번에서 시작하여 2, 3, 5, 4번 메뉴를 참조한 다음 답을 제시하였음을 나타내준다.

학생\메뉴	1	2	3	4	5	6	7	8	9	10
001	10	0	4	5	6	11	0	0	0	3
002	0	3	5	6	4	11	0	0	2	0

[그림 4] IMMEX에서 학생의 수행 과정을 숫자로 나타낸 표상의 예 (Vendlinski와 Stevens, 2000의 그림 2를 변형하였음.)

학생들의 정보의 검색 경로는 문제해결 전략으로 보고, Kohonen 네트워크를 이용하여 체계적으로 조직화할 수 있다. Kohonen 네트워크는 상관된 자료들을 공간적 지형으로 도해해 준다. 이 네트워크의 모든 구성마디(혹은 인공 신경원)는 다른 모든 마디와 연결되어 있고, 각 구성마디는 가까운 거리의 다른 구성 마디들과는 흥분적인, 먼 거리의 다른 구성 마디들과는 억제적인 상호작용을 일으킨다. 이 네트워크에 입력된 특정 패턴은 하나의 구성 마디에 최대의 반응을 일으키고, 이 구성 마디

는 다른 구성 마디들과의 흥분/억제의 상호작용 결과로 자기 조직화가 일어난다. 즉, 최대로 활성화된 마디를 중심으로 하여 모든 마디가 일종의 활성화 거품(bubble)을 형성하는데, 네트웍크의 작동이 계속되면서 구성 마디들 간의 상호작용의 정도가 변화되어 점차 이웃한 구성 마디들은 주어진 입력 패턴에 대해 유사한 반응을 보이는 독특한 공간적 지형을 형성하게 된다. 여러 입력패턴에 의한 학습과정이 진행되면 각 입력패턴에 대한 공간적 지형은 전체적으로 일정한 순서를 가진 지도를 형성하게 된다. 바로 이 이 출력 상태를 일종의 패턴화된 문제해결전략으로 볼 수 있다. 더욱 놀라운 것은 이처럼 유사성의 정도에 따른 자기조직화가 외부로부터의 어떤 수정이나 지시가 없이[4] 이루어진다는 점이다. 훈련 시간을 길지만 일단 학습이 일어나면 평가는 즉각적으로 이루어진다. 평가되어질 새로운 검색 경로가 입력되자마자 바로 이 입력에 가장 잘 대응되는 출력 마디가 활성화되기 때문이다.

IMMEX에서 Kohonen 네크워크를 사용하여 학생들의 수행을 분석하는 것의 장점 중의 하나는 이 분석의 결과가 교실 장면에서의 형성평가 도구로 활용될 여지가 많다는 점이다. 정보통신 기술을 이용한 대부분의 복잡한 수행평가는 평가 설계자가 완벽한 학생 모형 즉 초보자의 상태에서 전문가의 상태로의 전환 과정에 대한 충분한 이해를 바탕으로 한다. 하지만 문제의 유형에 따라 이 방식이 적절하지 않을 수 있는데, 그 대표적인 예는 최종 상태는 같더라도 다양한 중간 과정을 거칠 경우이다. IMMEX에서는 이런 발전 모형에 대한 사전 고려가 필요하지 않으며 전적으로 귀납적으로 만들어 낼 수 있다.

Hurst 등(1997)은 IMMEX에서 문제 해결에 실패할 경우, 그 원인이 학생이 잘못 이해하고 있어 특정한 검색 경로 패턴을 보임을 발견하였다. 이 정보는 학생들의 이해 방식을 추론하고 그에 적절한 교수법을 찾아내는데 즉각 활용될 수 있다. 또한 Vendlinski와 Stevens (2002)는 학생들이 메뉴를 사용하는 정도에 따라 과소, 적절, 및 과대의 세 유형으로 나누었다. 학생들이 첫 번째 문제에서 시작하여 여러 문제를 풀어가면서 이 전략이 어떻게 발전하는지를 Markov 연쇄분석을 이용하여 상태변환을 도해하였다. 이 결과는 교수법의 비교 등에 활용될 수 있다.

III. 선다형의 개선 방안

지금까지는 선다형의 문제점을 극복할 수 있는 구성형 시험 방식에 대해 살펴보았다. 하지만 앞에서도 강조한 것처럼 모든 상황에서 가장 잘 적용될 수 있는 최선의 평가 방식은 있을 수 없다. 시간이나 비용 등과 같은 평가에 따르는 여러 실질적 제약을 고려하고 절적한 방식을 찾는 것이 더 중요하다. 선다형의 장점은 여전히 활용될 필요가 있고, 실제로 개선되고 있다.

1. 적응형 컴퓨터화 검사

컴퓨터를 이용한 시험 방식 중 최근 가장 활발히 연구되는 분야는 적응형 컴퓨터화 검사이다 (예, Sands, Waters, & McBride, 1997; van der Linden & Glas, 2000; Wainer, 2000). 이 방식은 새로운 검사 이론인 문항 반응이론에 근거하고 있다. 문항 반응이론은 말 그대로 시험 문항을 하나하나 분석하여 문항 모수 즉 문항난이도, 문항변별도 및 문항 추측도를 안정적으로 추정하는 방법이다 (van der Linden & Hambleton, 1997). 이 방법은 일군의 집단에 대해 시험을 실시하고 각 피험자가 얻은 총점을 바탕으로 시험의 난이도, 변별도 및 추측도를 추정하는 방법과 대비된다. 고전 검사이론은 비교적 간단한 절차에 의해 문항 분석과 검사 분석을 할 수 있기는 하지만, 피험자 집단의 특성에 따라 그 분석 내용이 달라진다는 문제점이 있다. 하지만 문항 반응이론에 따른 분석을 통해 문항특성과 피험자 능력을 안정적으로 추정할 수 있기 때문에 현재 가장 널리 쓰이는 검사이론이다.

문항 반응이론의 발달과 함께 더 동질적인 문항으로 구성된 문제은행을 구축할 수 있게 되면서 적응형 컴퓨터화 검사가 등장하게 되었다. 이 방식은, 모든 수험자에게 똑같은 문항을 제시하는 대신에, 각 수험자의 능력 수준에 맞는 문항을 제시하여 시험을 보는 검사 방식이다. 이는 지필식에서는 불가능하지만 컴퓨터에서는 가능한 상호작용성(interactivity)을 활용하는 기법이다. 즉 수험자의 반응을 바탕으로 반응하여 컴퓨터가 그에 적절하게 문제를 제시하는 것이다. 이 검사 방식이 갖는 장점을 그물코를 비유하여 설명하자면, 하나의 크기를 가진 그물코로 고기를 잡는 대신, 여러 가지 다른 크기의 그물코를 가진 여러 겹으로 이루어진 그물을 던져 물고기를 잡는 동시에 그 크기대로 세분하여 모을 수 있게 하자는 것이다. 요컨대 수험자가 자신의 능력 수준에 맞는 문항을 충분히 풀 수 있게 되어 시험을 통한 변별력을 증가시킬 수 있다는 것이다.

적응형 컴퓨터화 검사는 현재 많은 수험자가 응시하는 시험에서 활발히 실시되고 있지만, 실행과 함께 새로운 문제점도 발견되고 있다. 그 중 몇 가지만 보면 다음과 같다. 우선 각 수험자에게 서로 다른 문항을 제시하지만, 실제로 이를 반복하다 보면 수험자 간에 같은 문항이 제시된다 (예, Mills & Steffan, 2000). 또한 만일 문제 은행에 충분한 수의 문항이 없이 반복해서 시험을 시행하다 보면, 검사 점수의 신뢰도가 떨어지는 문제점도 있다. 또한 충분한 수의 문항을 풀지 않은 수험자의 경우 점수를 어떻게 줄 것인지도 해결해야 할 난제이다.

이상은 실제로 적응형 컴퓨터화 검사를 연구하거나 실제로 시스템을 만드는 연구자들이 실제 장면에서 부딪친 문제들이다. 이들 외에도 적응형 컴퓨터화 검사를 실제 장면에서 사용할 때의 문제는 더 많다. 그 중 가장 큰 문제는 교사들이 실제 장면에서 이런 방식으로 시험을 보기가 어렵다는 것이다. 시험을 위해 문제은행에 엄청난 수의 문항을 미리 만들어 놓는 것은 그 자체로 교사에게 보통 일이 아니다. 게다가 각 문항의 난이도를 평정해야 한다고 하면 시험을 위해 이렇게 할 교사가 없으리라는 것을 짐작하기는 어렵지 않다. 일단 교사들이 부담을 느끼면 아무리 좋은 방식이나 절차가 있다고 하더라도 실제 장면에서는 쓰이지 않게 된다. 교사들은 자신들이 상황에 따라 쉽게 조작할 수 있는 방식으로 시험을 보기 원하며 이 때문에 지필식 선다형이 그 한계에도 불구하고 여전히 선호되고 있다.

적응형 컴퓨터화 검사와 관련하여 지적될 수 있는 또 다른 문제점은, 시험 방식이 기본적으로는 선다형이라는 것이다. 따라서 선다형 시험 방식이 갖는 단점을 그대로 떠안게 된다. 즉, 답지를 만들기가 어렵고, 고차 사고 능력을 제한하며, 추측의 가능성이 있다는 단점이 있다.

4) 신경망 모형 연구자들은 이를 unsupervised learning이라 부른다.

2. 다중 평가기법 (multiple evaluation)

선다형의 가장 큰 문제점의 하나는 추측에 의한 선택이다. 즉 답을 정확히 모르더라도 선택지의 구성적 특성이나 아니면 다른 추측전략을 사용하여 정답을 고를 수 있다. 이 문제점을 개선하면서 선다형이 갖고 있는 문제점을 극복하기 위한 방법이 다중 평가법이다. 이 방법은 수험자가 하나의 정답을 고르는 것이 아니라 제시된 모든 답지 하나하나에 대해 확률로 평가를 하는 방법이다.

이 방식은 오래 전에 Shuford와 그의 동료들에 의해 제안되었다 (Shuford, Albert, & Massengil, 1966). Shuford 등은 한 문제의 모든 답지에 대한 정답 가능성의 확신을 확률로 표시하도록 하되 이 확률의 합이 1이 되도록 하였다. 이 확률 분포로부터 정답에 대한 확신의 정도에 비례하여 점수화하는 방법은 로그 함수를 이용하는 방법뿐이다 (Shuford, et al., 136쪽). 그렇지만, 예를 들어 3개의 답지가 있을 때 .33보다 낮은 확률을 할당할 때처럼, 정답에 대한 확률 배점이 우연수준보다 낮을 경우 엄청나게 큰 감점이 이루어지게 된다. 이에 Shuford 등은 로그 변환을 사용하되 특정한 기준치 이하의 값은 -1로 변환시키는 방법을 제안하였다. 그 후 Dirkzwager (1997)는 추측 허용치 (tolerance-for-guessing parameter) 개념을 도입하여 이 문제를 수학적으로 더 정교한 방식으로 해결하였다.

점수화 공식과 더불어 다중 평가 기법에서 핵심적으로 중요한 개념은 실재론(realism)이다. 만일 어떤 학생이 정답에 대한 확신이 .75라 하자. 그런데 실제로 .75라 반응한 문제 가운데 75%가 실제 정답이었다면 이 학생은 완전히 실재적이다. 이와는 달리 만일 정답에 대한 확신이 .75인데 실제 정답률이 이보다 높거나 낮게 반응하는 학생은 과잉확신하거나 과소확신하고 있음을 나타내준다. 이들에 대해서는 "당신은 지금 자신의 지식에 대해 과신하고 있거나 아니면 지나치게 추측하고 있습니다" 혹은 "당신은 당신이 생각하는 것보다 더 많이 알고 있습니다. 더 확신을 높이고 과감해지십시오" 등과 같은 피드백을 시험을 보는 중간 중간에 제시해 줄 수 있다.

시험이 끝나고 나면 최종 점수와 함께 그들이 얻으려했던 점수, 그리고 위에서 제시했던 피드백과 함께 실재론 점수를 제시하여 이후에 같은 방식으로 시험을 볼 때 더 실재적으로 볼 수 있도록 유도한다. Dirkzwager (1996)는 11세를 대상으로하여 이 방식이 성공적으로 적용될 수 있음을 보여주었고, Holmes(2002)는 이 방식이 전통적 선다형보다 신뢰도나 타당도가 높음을 보여주었다.

3. 컴퓨터를 이용한 변형 선다형 시험 방식

CMMT(computerized modified multiple-choice testing) 시스템은 선다형시험이지만 마치 단답형을 풀도록 한 다음 선다형 답지를 이용하여 최종 반응을 하게 하는 시험방식이다 (Park, in press). 이 방식은 앞에서 지적된 선다형의 문제점인, 사고를 수동적으로 만들 수 있으며 또 답지를 활용한 추측을 어느 정도 해결할 수 있다. 이를 몇 개의 단계로 나누어 설명하자면 다음과 같다 (그림 5 참조). 수험자들이 화면 하단의 문항 번호에 마우스를 갖다 대면 문제 제시화면에 답지가 없이 문제만을 제시하되, 정해진 시간 동안에는 문제를 얼마든지 볼 수 있다. 특정한 문제 (예를 들면 4번)에 대해 수험자가 답을 스스로 찾았다고 생각하면, 마우스를 클릭하여 답지를 제시하도록 요구한다. 그러면 답지가 정해진 시간만큼 (답지의 길이나 이해도에 따라 달라지는데, 4번 문제의 경우 4초) 짧게 제시되는데, 수험자는 그 시간 내에서만 반응할 수 있다. 그 시간이 지나면 그 답지와 함께 문제 자체가 화면에서 사라져 더 이상 반응을 할 수 없게 된다. 이 방식의 핵심은 수험자가 기본적으로 단답식처럼 문제를 풀고, 마지막 순간에 답지를 이용하여 자신이 생각한 내용을 골라 반응하게 하자는 것이다.

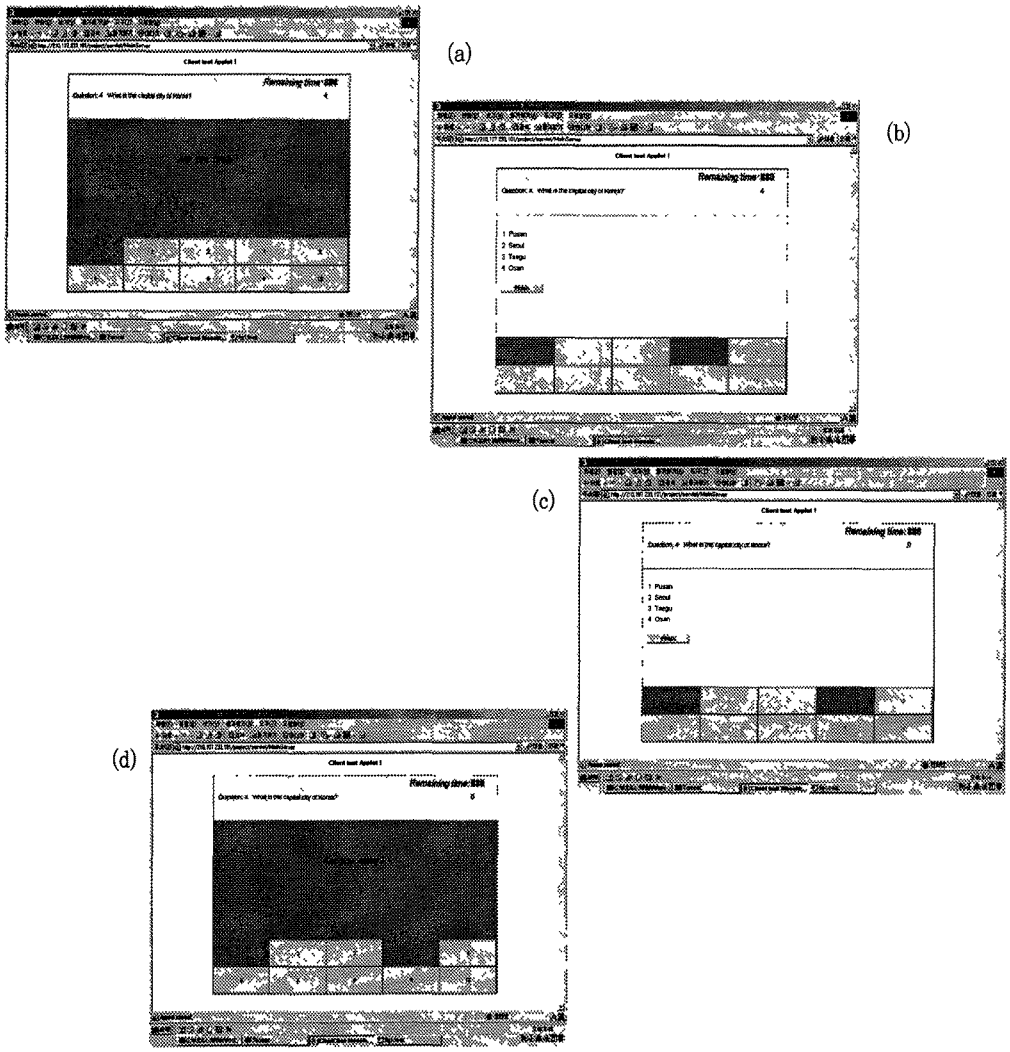
이런 시험 방식을 통해 기대되는 효과는 우선 단답형에서처럼 선다형에서 보다 인출의 난이도를 증가시킬 수 있다는 것이다 (예, Bjork, 1994). 실제로 Park (in press)은 이 방식으로 중간 시험을 보았을 때 전통적 선다형으로 중간 시험을 보았을 때보다 최종 시험에서 더 높은 점수를 얻음을 발견하였다. Park과 Choi (2005)는 컴퓨터로 보는 시험을 집에서 보게 했을 때에도 CMMT방식으로 볼 때가 전통적 선다형 방식으로 볼 때보다 최종 시험에서 더 높은 점수를 받는 결과를 얻었다. 이상의 결과는 CMMT가, 전통적 선다형보다, 인출의 난이도를 어렵게 하여 나중에 다시 기억해 낼 때 기억을 더 고양시켰음을 의미한다.

CMMT 방식의 또 다른 장점은 답지를 나중에 짧게 제시하기 때문에 답지만 가지고 추측하여 선택하는 것을 어느 정도 배제할 수 있다. 실제로 Park(2005)은 컴퓨터화 시험을 두 가지 형식으로 보되 시험 직후에 피드백을 제시하지 않았다. 4일 후 예고 없이 동일한 문제를 지필식 단답형으로 다시 풀게 하였다. 이 연구의 종속측정치는 컴퓨터로 보았을 때 맞은 문제를 단답식으로 보았을 때 다시 맞출 확률, 즉 P(C2|C1)이었다. CMMT 방식으로 본 집단의 경우 이 조건부 확률은 .67이었고 전통적 선다형으로 본 집단은 .57이었는데 이 차이는 통계적으로 유의미하였다.

이 밖에 반응 자체는 선다형처럼 답지를 고르는 것이기에 채점이 쉽고 객관성을 유지할 수 있다는 점, 그리고 답지를 만드는 일이 상대적으로 쉬워져 출제가 용이해진다는 점도 장점이라 할 수 있다.

IV. 전망과 과제

교육의 질을 높이기 위한 노력은 그야말로 전방위적으로 이루어지고 있다. 인지심리학에서 발견된 주요 내용들, 예를 들면 장기 기억에 지식이 어떤 식으로 체제화되어 있으며, 한 영역에 대한 전문성은 어떤 식으로 발전하는지, 그리고 선행 지식은 어떻게 영향을 주며 전이는 언제 잘 이루어지는 지 등에 대한 연구 결과를 바탕으로 교수 학습 설계가 이루어지고 있다 (예, Bransford, Bown, & Cocking, 2000). 또한 사회적 차원을 고려하는 교사의 역할 및 수업 방식에 대한 연구도 활발하다 (예, Brown & Campione, 1994).교수 및 학습 영역에서의 이런 변화는 평가 방식에도 큰 변화를 일으키고 있다. 평가의 초점은 그 어느 때보다도 단순한 사실에 대한 기억보다는 실제적이고 복잡한 추리 과정을 추적할 수 있는 방향에 맞추어지고 있다. 소위 수행평가로 통칭되는 구성형이 각광을 받게 된 것은 이런 배경에서이다.



[그림 5] CMMT의 작동 예. (a) 1 번 문제는 풀었기 때문에 1번 칸은 사라졌다. 마우스의 위치는 4번 칸에 놓여있기 때문에 4번 문제인 "What is the capital of Korea?"가 문제 제시화면에 나타나있다. 오른쪽 위의 숫자 4 (화면에는 붉은 색으로 나타남)는 이 문제에 대해 반응할 시간이 4초로 정해져 있음을 나타내준다. (b) 수험자가 4번 칸에 마우스를 클릭하면 답지가 제시되면서 시간이 줄어들기 시작한다. (c) 수험자가 자신이 생각한 답을 선택하면 선택된 답지가 붉은 색으로 변화된다. (d) 4초가 지나 선택지가 사라지고 수험자는 더 이상 반응을 할 수 없다.

그렇다고 해서 구성형의 도입이 모든 평가 문제의 해결책이라고 생각해서는 안된다. 왜냐하면 선다형과 구성형은 지적인 능력의 서로 다른 측면을 다루고 있기 때문이다.

구성형은 산출과정을 중시하는 반면, 선다형은 차이의 구별에 비중을 둔다. 이와 함께 고려해야 할 변인은 비용이다. 구성형이 선다형에 비해 드는 추가 비용은 엄청나기 때문에 대부분의 경우 선다형이 선호된다. 더욱이 선발을 위한 시험의 경우 구성형이 선다형보다 더 낫다는 증거도 없다.

그럼에도 불구하고 선다형일변도의 시험을 탈피하는 것은 필요하고 바람직하다. 상당수의 학생들은 학습 자체보다는 좋은 평가에 더 큰 관심을 가지고 있고, 따라서 더 열심히 공부하기보다는 선다형 시험을 잘 보는 방법을 훈련하기도 한다. 선다형을 탈피해야 하는 더 중요한 이유는 형성 평가도구로서는 구성형이 선다형보다 더 효과적이라는 점 때문이다.

이런 맥락에서 위에서 소개된 컴퓨터를 활용한 여러 혁신적인 시험 방식은 교육의 질을 높이는데 중요한 역할을 할 것으로 기

대된다. 구성형 시험의 채점과 관련된 비용문제를 해결할 수 있을 뿐만 아니라, 선다형에서도 변별 이외의 능력을 측정할 수 있도록 해주기 때문이다.

그렇지만 아직도 이들 시험 방식이 교육 현장에 적용되어 교육의 질을 높이는 데 실질적으로 기여하기 위해서는 해결해야 할 난제가 많이 있다. 그 하나는 이들 시스템이 아직은 비싸거나 사용하기가 그리 쉽지 않다는 점이다. 프로그램에 대한 최소한의 이해와 데이터 베이스에 대한 기본적인 지식과 조작 능력이 필요하기 때문이다. 이보다 더 큰 문제는 교사들의 태도이다. 대부분의 경우 교사들은 변화를 싫어하는데, 평가의 경우도 예외가 아니다 (예, Marzano, 2000). 위에서 언급된 기술적인 문제는 약간의 노력으로 극복될 수 있는데 그 열쇠는 교사의 동기이다. 새로운 기술을 경쟁자로서가 아니라 교육의 질을 향상시킬 수 있는 도구로 받아들이고 이를 실제 장면에 활용하기 위해서는 교사들의 적극성이 절대적으로 필요하다. 다른 모든 교육 개혁과 마찬가지로, 평가에서도 교사의 참여가 없이는 성공할 가능성이 없다고 해도 과언이 아니다. 일단 교사들이 참여하게 되면, 사용이 간편하면서도 그 효과를 확인할 수 있는 실제적 경험이 필요하다. 따라서 평가의 중요성에 대한 이해를 고양시키는 한편 현장에서 활용 가능한 프로그램 사용법에 대한 단기 연수가 효과적일 것으로 예상된다.

스스로 정보 통신 분야의 강국으로 일컫지만, 우리의 e-testing은 그리 자랑할 만한 수준이 아니다. 아직까지는 교육 분야에서 게임처럼 세계에 내놓을 만한 새로운 기술은 눈에 띄지 않기 때문이다. 앞에서 소개된 대부분의 기술은 주로 미국을 중심으로 개발되고 있고, 네덜란드, 영국, 호주, 중국, 그리고 대만 등도 나름대로의 시스템을 개발하고 구축하기 위해 박차를 가하고 있다. 특히 미국은 일찍부터 평가 이론가, 인지 심리학자, 교육공학자, 그리고 컴퓨터 공학자들이 서로의 전문성을 활용하여 새로운 기술을 개발해왔다. 물론 이에 대한 전국가적인 지원이 있었음을 간과할 수 없다.

e-testing은 아직 새로운 분야이지만 e-learning의 활성화와 발맞추어 비약적으로 발전할 분야가 될 것이다. 이미 늦은 감이 없지 않지만, 지금이라도 창의적이고 실용적인 기법 개발을 위한 정책적인 지원을 토대로 연구자들의 분발과 협력이 그 어느 때보다 필요한 시점이라 하겠다. 그리고 학생들은 물론 교사들의 적극적인 참여를 이끌어 낼 때 비로소 e-testing이 교육의 질을 높이는 중요한 도구로 사용될 수 있을 것이다.

참고 문헌

Ager, T. A. (1990). From interactive instruction to interactive testing. In R. Freedle, (Ed.),Artificial intelligence and the future of testing. (pp. 21-52). Hillsdale, NJ: Lawrence Erlbaum Associates.

Armrein, A.L., & Berlinger, D.C. (2002). High-stake testing, uncertainty, and student learning. Educational Policy Analysis Archives, 10(18). http://epaa.asu.edu/v10n18.html.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences* **4**(2). 167-207.

Bennett, R. E. (2001). How the internet will help large-scale assessment revent itself. *Educational Policy Analysis Archives.* 9(5). http://epaa.asu.edu/epaa/v9n5.html.

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *The Journal of Technology, Learning, and Assessment* 1(1).

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human being. In J. Metcalfe & A. P. Shimamura (Eds.) *Metacognition.* MIT Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-73.

Bransford, J.D., Brown, A.L., & Cocking, R.R. (2000). *How people learn.* National Academy Press.

Brown, A .L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice.* Cambridge, MA. MIT Press.

Bunderson, C. V., Inouye, D. K., & Olson, J. B.(1989). The four generations of computerized educational measurement. In R. I. Linn (Ed.). *Educational measurement (3rd ed.)* NY: American council of education.

Burstein, J., Chodorow, M., Leacock, C.(2003). Criterion online essay evaluation: an application for Automated evaluation of student essays. *Proceeding of the Fifteenth Annual Conference on Innovative Applications of Artificial Intellignence, Acapulco, Mexico.*

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE Stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems* 18 (1). 32-39.

Burton, J. K., Moore, M. M. & Magliaro, S. G.(1996). Behaviorism and instructional technology. In D. H. Jonasson (Ed.) *Handbook of research for Education Research and Communications,* (pp. 46-73). New York: Simon & Schuster Macmillan.

Chung, G. K. W. K., Baker, E. L., & Cheak, A. M. (2002). *Knowledge mapper authoring System Prototype.* CSE Technical Report 573. University of California, Los Angeles.

Dirkzwager, A. (1996). Testing with personal probabilities: eleven year olds can correctly estimate their personal probabilities. *Educational and psychological measurement.* 56. 957-971.

Dirkzwager, A. (1997). *A Bayesian testing paradigm: multiple evaluation, a feasible alternative for multiple choice.*

Earl, M. L. (2003). *Assessment as learning.* Corwin Press, Thousand Oaks: CA.

Foos, P.W., & Fisher, R.P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology, 80(2),* 179-183.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist* 39. 193-202.

Frederiksen, J. R., & White, B. Y. (1989). Intelligent tutors as intelligent testers. In N.Frederiksen, R.Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition.* (pp. 1-25). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glover, J. (1989). The "testing" phenomenon: Not gone, but nearly forgotton. *Journal of Education Psychology,* Vol. 81 (3), 392-399.

Goldmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Education Psychology,* Vol. 83 (1), 88-96.

Greer, J., & McCalla, G. (1994). *Student modelling: The key to individualized knowledge-based instruction.* Springer-Verlag.

Harp, S. A., Samad, T., & Villano, M. (1995). Modeling student knowledge with self-organizing feature maps. *IEEE Transactions on Systems, Man, and Cybernetics 25(5).* 727-737.

Herl, H. E., & Baker, E. L., Niemi, D.(1996). Construct validation of an approach to modeling cognitive structure of U. S. history knowledge. *The Journal of Education Research. 89(4),* 206-218.

Herl, H. E., O'Neil Jr. H. F., Chung, G. K. W. K., & Schacter, J. (1999). Relability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior,* 15. 315-333.

Holmes, P. (2002). *Multiple evaluation versus multiple choice as testing paradigm.* Unpublished doctoral dissertation. University of Twente.

Hurst, K., Casillas, A., & Stevens, R. H. (1997). *Exploring the dynamics of complex problem-solving with artificial neural network-based assessment systems.* CSE Technical Report 444. National Center of Research on Evaluation, Standards, and Student Testing(CRESST). University of California, Los Angeles, CA.

Jonassen, D. H., Tessmer, M., & Hannum, W. H. (1999). *Task analysis methods for instructional design.* LEA: Mahwah. New Jersey.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes 25.* 259-284.

Lesh, R. & Kelly, A. E. (1996). A constructivist model for redesigning AI tutors in mathematics. In J. M. Laborde (Ed.) *Intelligent learning environments. The case of geometry* (pp. 134-156). New York, NY Grenoble: Springer.

Linn, R. (2000). Assessment and accountability. *Educational Researcher, 29*(2), 4-16.

Marzano, (2000). Transforming classroom development. ASCD, Alexandria:VA.

Mills, C. N., Steffen, M. (2000). The GRE computer adaptive test' Operational issues. In W. J. van der Linden & C. A. W. Glas(Eds.). *Computer adaptive testing. Theory and practice (pp.75-99).* Norwell, MA: Kluwer Academic.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction.* 5, 253-292.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis with implications for designing simulation-based performance assessment. *Computers in Human Behavior, 15.* 335-374.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47(1),* 238-243.

Page, E., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76(3),* 561-565.

Park, J. (2004). *Testing phenomenon in the computerized modified multiple-choice testing system.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, April 15.

Park, J. (2005). *Higher retention of the correct answer in a new computerized testing system.* Manuscript submitted for publication.

Park, J. (in press). Learning in a new computerized testing system. *Journal of Educational Psychology.*

Park, J., Choi, B. (2005). *Learning at home using a new computerized testing system.* Manuscript submitted for publication.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know.* National Academic Press.

Rudner, M. (2002). *Measurement decision theory.* Manuscript submitted for publication.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning, and Assessment, 1(2).*

Ruiz-Primo, M. A., Shavelson, R. J., Schultz, S. E. (1997). *On the validity of concept map-base assessment interpretations: an experiment testing the assumption of hierarchical concept maps in science.* CSE Technical Report 455. Stanford University.

Ruiz-Primo, M. A., Schultz, S., Li, M., & Shavelson, R. J. (1999). *On the cognitive validity of interpretations of scores from alternative concept mapping techniques.* CSE Technical Report 503. Stanford University.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Schacter, J., Herl, H. E., Chung, G. K. W. K., Dennis, R. A., O'Neil Jr., H. F. (1999). Computer-based performance assessments' a solution to the narrow measurement and reporting of problem-solving. *Computers in Human Behavior, 15.* 403-418.

Shuford, E. H. Jr., Albert, A., & Massengil, H. E. (1966). Admissible probability measurement procedures. *Psychometrika. 31.* 125-145.

Stevens, R., Ikeda, J., Casillas, A., Palacio-Cayetano, J., Clyman, S. (1999). Artificial Neural Network-Based Performance Assessments. *Computers in Human Behavior.* 15. 295-313.

Thompson, C. (1999). New word order: The attack of the incredible grading machine. *Ligua Franca July/August,* 28-37.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York Springer.

van der Linden, W. J., & Glas, C. A. W. (2000). *Computer adaptive testing: Theory and practice.* Norwell, MA: Kluwer Academic.

Vendlinski, T., & Stevens, R. (2000). *The Use of Artificial Neural Nets (ANN) to Help Evaluate Student Problem Solving Strategies. Proceedings of the Fourth International Conference of the Learning Sciences.* Mahwah, NJ. 108-114.

Vendlinski, T., & Stevens, R. (2002). A Markov Model Analysis of Problem-Solving Progress and Transfer. *Journal of Technology, Learning, and Assessment 1(3).*

Wainer, H. (2000). *Computerized adaptive testing' A primer (2nd ed.).* Hillsdale, NJ: Erlbaum.